FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2019/2020

Trabajo de Fin de Máster

TITULO: *Predicción de abandono de clientes en una empresa de telecomunicaciones*

***Alumno*:** Federico Alegre

***Tutor*:** Ramón Alberto Carrasco

Julio de 2020

UNIVERSIDAD COMPLUTENSE
MADRID

A mi familia y amigos.

Resumen

El abandono de clientes refleja el comportamiento de una compañía para con estos clientes y el comportamiento de la competencia. Cada vez es más relevante conocer las causas que provocan que un cliente no compre más productos o no continúe un contrato de servicios. Poder predecir con cierta confianza si un cliente abandonará la compañía antes que suceda puede generar grandes ahorros para una organización y permitirle desarrollar mejores prácticas y procesos en el futuro. Este trabajo busca predecir este fenómeno a través de diferentes algoritmos de *machine learning* para luego priorizar las acciones de retención de clientes mediante la segmentación de los mismos.

Palabras clave: Abandono de cliente, predicción de abandono, machine learning, variables, datos, modelos, clientes, empresa

Abstract

Customer churn reflects the behavior of a company towards those customers and also the behavior of the competition. It is becoming increasingly relevant to know the causes that result in the customer not buying more products or withdrawing from a service agreement. Being able to predict, with certain reliability, if a customer is going to churn before it happens could derive in big savings for company and allow it to develop better business practices and processes in the future. This paper attempts to predict this phenomenon through the use of different machine learning algorithms in order to prioritize the company retention actions based on customer segmentation.

Keywords: Customer churn, churn prediction, machine learning, variables, data, models, customers, company

Índice

Lista de figuras

Lista de tablas

1 Introducción

1.1 Contexto

¿Qué es lo que motiva a las empresas a existir? O, mejor dicho, ¿Cuál es su objetivo? Ganar dinero sería la respuesta más comentada probablemente.

¿Cómo pueden cumplir su objetivo? A través de la provisión de un servicio y/o producto a alguien que pague por ello, es decir, clientes.

¿Qué se puede hacer con estos "clientes"? Desde una perspectiva de marketing, las acciones a tomar son: adquirir nuevos consumidores, retener los clientes existentes y desarrollarlos, o dicho de otra manera, lograr que el mismo compre más de un producto o de otro que la empresa provea.

¿Qué sucede en un contexto de alta competencia y con baja tasa de crecimiento? En una situación de madurez de mercado ganar dinero a través de la adquisición de nuevos clientes se transforma en una tarea cada vez más ardua y dificultosa.

¿Es sostenible en el tiempo avocar todos los recursos en la adquisición de clientes? Evidentemente no. Aunque para algunas industrias pareciera que la cantidad de personas dispuestas a contratar o comprar es infinita, esta no lo es. Cada organización busca una cuota de mercado mayor, lo que provoca que sea más dificultoso ganar nuevos clientes.

¿Qué puede hacer una empresa en este caso? Dedicar una mayor cantidad de recursos a la retención y desarrollo de los clientes existentes. A su vez, con un correcto trabajo en estos y un análisis exhaustivo de los clientes que mayor valor generan, es posible extraer conocimientos que permitan emplear formas de adquisición de clientes más precisas y menos costosas.

¿Cómo puede materializarse? Una vez que la empresa ha identificado su público objetivo o sus segmentos de clientes, al estudiarlos y comprender cuáles son sus características, el uso que realizan de sus productos, los momentos de compra, las influencias que tienen, cómo viven y un número casi infinito de datos sobre estos, esta puede utilizar todo este conocimiento adquirido para buscar más personas como los clientes que más valor le proveen. Por consiguiente, es posible garantizar la supervivencia de la organización con las acciones comerciales adecuadas y una correcta gestión de los clientes y recursos.

En este trabajo se procederá a realizar un estudio para poder predecir la probabilidad con la que un cliente de determinada compañía deja de contratar sus servicios.

La importancia del estudio del abandono de un cliente, o "churn rate", puede verse reflejada en diferentes aspectos:

- El coste de mantener a un cliente es inferior al coste de adquirir un nuevo cliente. Diferentes autores/as han determinado una relación de diferentes magnitudes, pero con la misma idea de base:
    - "Adquirir nuevos clientes cuesta entre cinco y seis veces más que retener a clientes existentes" (Verbeke et al., 2012)

    - “El coste de adquisición de un cliente es más grande que el coste de retención del mismo, en algunos casos puede ser hasta 20 veces más caro” (Vafeiadis et al., 2015)
    - “Conservar un cliente existente es cinco veces más barato que atraer uno nuevo” (McIlroy & Barnett, 2000)
    - “Dependiendo en que estudio uno se base, o en qué industria se esté, adquirir un nuevo cliente es entre cinco y 25 veces más costoso que retener uno existente” (Gallo, s. f.)
- La relación a largo plazo con clientes es más rentable:
    - “Clientes de largo plazo generan ganancias más altas, con una tendencia a ser menos sensibles a las acciones de marketing de la competencia y son menos costosos de retener” (Verbeke et al., 2012)
- Aumenta la rentabilidad de la compañía:
    - “Mejorar la retención de clientes contribuyó en la reducción de la tasa de abandono de 20% a 10% anual, que permitió ahorrar unos 25 millones de libras a la operadora Orange” (Aydin & Özer, 2005)
- Establecer prioridades en cuanto a las acciones del área de marketing:
    - “Adquirir un cliente nuevo o recuperar un cliente perdido es más costoso que retener a un cliente existente. Sin embargo, actualmente los presupuestos de marketing están más dirigidos a la adquisición de cliente (60%) que a la retención (40%)” (IE Catedra de Fidelización et al., 2018)

El contexto mundial ha cambiado en sus diferentes pilares. En lo que respecta a los negocios y las empresas, los clientes cada vez son más exigentes en cuanto a calidad, precio, prestaciones, servicios derivados de la venta o del servicio principal, y acciones de las compañías más allá de su actividad principal, como el involucramiento con cuestiones climáticas y sociales.

En la mayoría de los mercados hace bastante tiempo se transitó de una situación en el cual la demanda era mayor que la oferta, para convertirse en un mercado de oferta abundante y clientes difíciles de retener.

En este contexto es donde conocer a los clientes, entender sus necesidades, generar una relación fluida con ellos parece el camino hacia la sostenibilidad de una compañía a largo plazo.

Conceptos como la satisfacción, la lealtad, el “churn rate” (tasa de cancelación), “Customer Lifetime Value” (V. Kumar, 2018), “Customer Influencer Value” (V. Kumar, 2018), “Customer Referral Value” (V. Kumar, 2018), “Customer Knowledge Value” (V. Kumar, 2018) y demás términos utilizados hoy en día, cada vez cobran un mayor peso a la hora de tomar una decisión de compra para las personas.

Es por ello que determinar con un alto grado de certeza si un cliente muestra una tendencia a abandonar, con un tiempo necesario para tomar acciones preventivas, parece ser una buena medida para establecer una base sólida que permita sostener a una organización y permitirle mejorar sus ofertas al resto de la sociedad.

1.2 Concepto y Origen

¿Qué implica retener a un cliente? ¿Cuándo se considera que el cliente ha abandonado la empresa? ¿Qué datos podemos utilizar para predecir un fenómeno que ni el propio cliente puede saberlo? ¿Cómo anticiparse a esto? ¿Qué medidas pueden resultar efectivas? ¿Todos los clientes son igualmente importantes? ¿Qué variables ayudan a predecir el abandono?

Las interrogantes anteriores muestran la necesidad de comenzar definiendo qué es el abandono de cliente o "churn".

Los clientes que abandonan a una empresa o proveedor son aquellos que o bien dejan de adquirir sus servicios o productos, o cesan en forma voluntaria o involuntaria su relación contractual.

La tasa de abandono de clientes (Churn Rate) se puede obtener de la siguiente forma:

$$Churn\ Rate = \frac{Clientes\ que\ permanecen\ en\ la\ compañía\ a\ final\ del\ periodo}{Clientes\ al\ inicio\ del\ periodo} \times 100$$

Dicho periodo al que hace referencia la fórmula puede ser mensual, bimensual, trimestral, semestral, anual o por el tiempo que la empresa decida realizar dicha evaluación. Este lapso de tiempo será determinado por la velocidad de cambio de la base de clientes que tenga la compañía o el valor que utilice la industria o sector en la que esta se encuentra.

En el sector de telecomunicaciones es habitual utilizar un periodo mensual o trimestral; en el sector bancario y de seguros un lapso de 12 meses; en el sector de retail puede utilizarse de una semana, un mes u otro periodo dependiendo de la frecuencia de compra de un cliente promedio.

El cálculo de abandono de un cliente difiere enormemente si la relación cliente-proveedor se encuentra reflejada en un contrato o no.

Para el cálculo de abandono de relaciones no contractuales debe estudiarse el tiempo habitual de recompra de un cliente. Para ello, se establece un periodo de calibración y un periodo de estudio en el cual se determina si un cliente comprará o no comprará en un plazo determinado posterior a una compra. Si el cliente realiza una transacción con el proveedor, se lo determina como no abandono; si el cliente no realiza una transacción en dicho periodo de estudio, se lo considera abandono.

Para el caso de relaciones contractuales es un escenario más simple: la relación se acaba cuando alguna de las dos partes finaliza dicha relación de forma explícita o no se genera una renovación del contrato una vez expirado el mismo.

Existen diferentes tipos de "churn", como proponen Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr en su estudio (Shaaban et al., 2012):

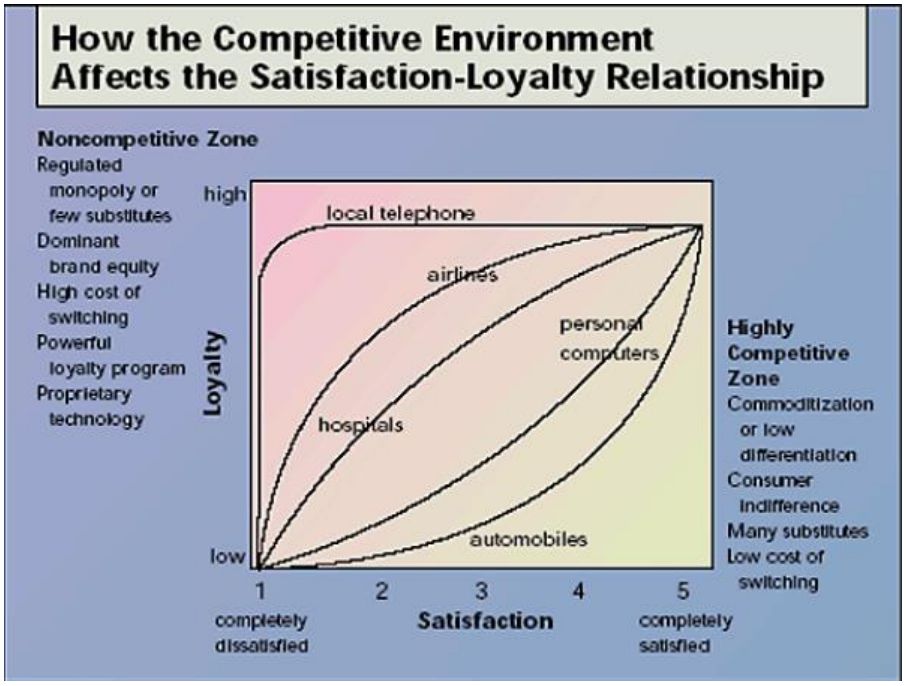
- Según quien decide finalizar la relación:
    - Voluntario: el cliente.

- Involuntario: la empresa puede determinar finalizar la relación, siendo las causas más frecuentes las de fraude, uso inadecuado del servicio, falta o problemas en los pagos o decisión comercial.

- Dentro del abandono voluntario, puede encontrarse una subdivisión:
  - Accidental: no planeado por el cliente, sino a causa de algo que ocurrió en su vida. Por ejemplo: cambio de domicilio, cambio en condiciones financieras, entre otras.
  - Deliberado: planeado por el cliente, pudiendo estar originado por causas: tecnológicas, económicas, factores en la calidad del servicio, otro tipo de conveniencias. Este tipo de abandono suele estar influido por acciones de la competencia.

Se podría pensar que el abandono, o en su contrapartida, la retención se ve directamente afectada por la satisfacción del cliente con el servicio provisto o el producto adquirido, y que esto conlleva a la lealtad del mismo con el proveedor.
Pero, como se muestra a continuación, existe un estudio realizado sobre la relación entre satisfacción-lealtad en distintos entornos:

Figura 1. ¿Cómo el contexto competitivo afecta la relación satisfacción-lealtad?



(Jones & Sasser, s. f.)

Como puede observarse en la Figura 1, dependiendo del negocio en el cual se encuentre una compañía, el efecto que tiene la satisfacción del cliente se relaciona en forma diferente con la lealtad hacia la misma.

Diferentes variantes pueden conjugar un espacio distinto en cuanto a la aplicación de un modelo de predicción de abandono, como, por ejemplo: velocidad demandada de acción del negocio, facilidad de abandono del cliente, coste de oportunidad del cliente si abandona, barreras de salida, precio del bien o servicio, tipo del bien o servicio, categoría y demás variables.

Esta realidad presenta una alta complejidad debido a la cantidad de factores que se interrelacionan para dar lugar a que una persona decida finalizar su relación con un proveedor y buscar satisfacer la necesidad en cuestión en otro sitio.

1.3 Relevancia y justificación

En un mundo globalizado, en el cual las distancias ya no son tan extensas como lo eran hace tiempo, en el cual el comercio, la oferta, la demanda, el consumo y la competencia avanza cada vez con mayor velocidad, la retención de clientes es un factor clave para que la empresa subsista, pueda crecer y desarrollarse.

Nos encontramos en una situación inédita con respecto al siglo pasado. De la noche a la mañana puede surgir un nuevo servicio, una nueva compañía o un nuevo producto que puede cambiar al mundo y hacer peligrar la posición de las empresas más poderosas.

Esta nueva realidad que cambia constantemente ha impactado en forma trasversal a todos los aspectos de la vida, y el mercado no es la excepción.

En diferentes industrias y sectores, en los cuales comienzan a emerger signos de reducción de la velocidad de crecimiento o madurez, consolidar la posición competitiva y asegurar los activos que producen beneficios parece ser una de las únicas medidas que tienen a su alcance las organizaciones.

Uno de los activos más importantes, sino el más relevante, de las empresas son sus clientes. Son estas personas, físicas o jurídicas, que permiten a una entidad crecer y prosperar. Una compañía puede tener más o menos recursos que otra, puede competir desde una posición desventajosa en comparación, pero una empresa sin clientes no posee ningún tipo de justificación de existencia.

Es por esto mismo que, desde el punto de vista del marketing, se ha establecido la importancia que debe tomar la cartera de clientes, no solo para elevar el valor que propone este área o sector de la compañía, sino porque en la clientela está la clave de cada acción y paso que da una compañía. Establecer acciones centradas en el cliente, generar una relación con este, entendiendo sus necesidades y proponiendo formas de satisfacerlas es uno de los grandes desafíos con los que los negocios se enfrentan día a día.

Una alta competencia en un sector implica, en la mayoría de los casos, una transferencia de poder hacia la demanda, es decir los clientes, quienes tienen un abanico más amplio de ofertas para juzgar, comparar y elegir. Por lo que conocer qué es lo que motiva a un cliente a escoger una empresa por sobre otra, o generar una relación a largo plazo con alguna de ellas constituye un fenómeno que debería ser de interés principal para las compañías.

Entonces, conocer con qué probabilidad un cliente puede abandonar una empresa, desde dejar de consumir sus productos o no continuar con la suscripción a un servicio, parece un hecho de vital relevancia para cualquier compañía actual.

La satisfacción del cliente puede que no lo sea todo, la lealtad a una compañía o marca no significa la continuidad como cliente ni garantiza que la empresa tendrá su futuro asegurado. Es por ello que detectar si un cliente puede abandonar la organización, y con qué probabilidad, puede ser un diferencial sustancial.

Existe un amplio número de estudios realizados sobre la probabilidad de abandono, pero no todos estos proponen una forma práctica de abordarlo desde una perspectiva de marketing. Además de conocer la probabilidad de abandono de un cliente, es importante descubrir de qué forma es posible retenerlo y si es el objetivo de la compañía ese segmento en particular.

No todos los clientes son de interés para una empresa, no todos los segmentos son rentables ni son el foco estratégico de atención de la organización.

A su vez, una acción de retención puede ser efectiva para un cliente, pero en otro puede incluso aumentar la probabilidad de abandono.

Conocer qué motiva a un cliente, o a un grupo de ellos, y cómo utilizarlo en forma efectiva, con los recursos escasos de la organización, es un concepto clave en un estado de competencia constante.

Por ende, utilizar los recursos de forma adecuada y escoger qué clientes retener y cuáles dejar ir es un factor diferencial, que puede implicar la desaparición o la prosperidad a largo plazo de una compañía.

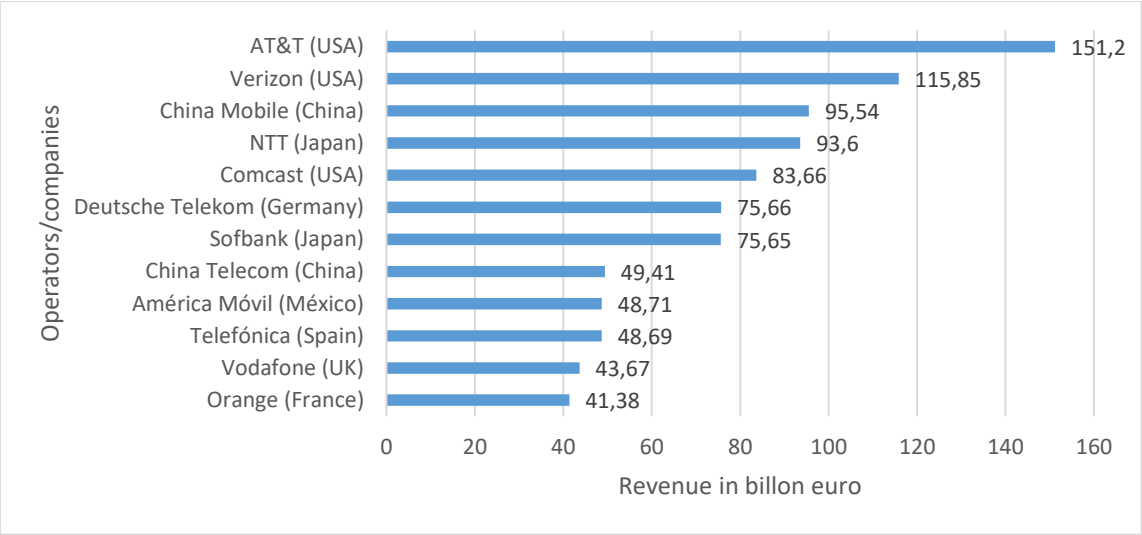
1.3.1 Números de la industria de telecomunicaciones

A raíz que los datos que se utilizan en este trabajo en cuanto a la predicción de probabilidad de abandono corresponden a una empresa del sector de telecomunicaciones, a continuación, se presentan algunos valores de dicho sector para comprender su magnitud e impacto.

1.3.1.1 Valores mundiales

A continuación, se muestran las 12 compañías que mayores ingresos han tenido en 2018 del sector de telecomunicaciones:

Figura 2. Ingresos de operadores de telecomunicaciones a nivel mundial en 2018

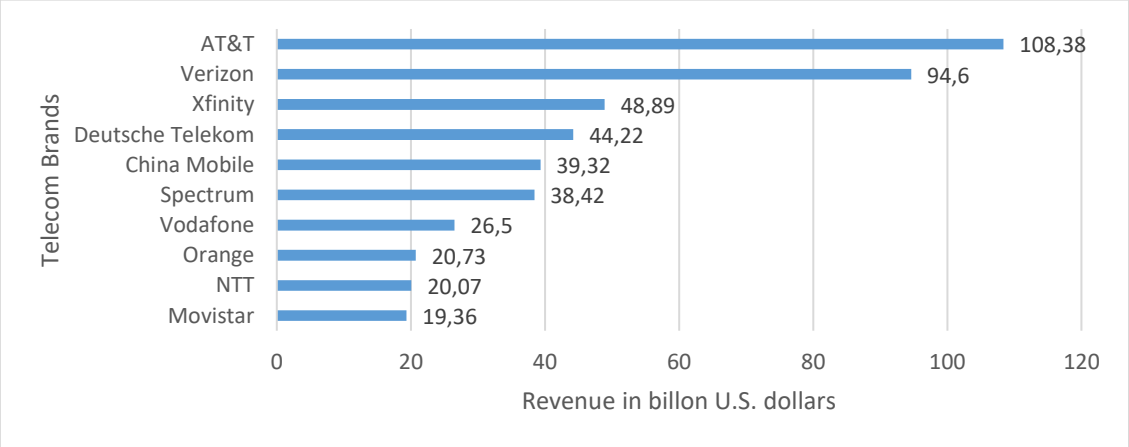


Fuente: Statista.

Se evidencia en la Figura 2 que el mercado de Estados Unidos es uno de los que más ingresos genera a nivel mundial.

A nivel de valor de marca se observa, en la Figura 3, que AT&T y Verizon lideran el ranking:

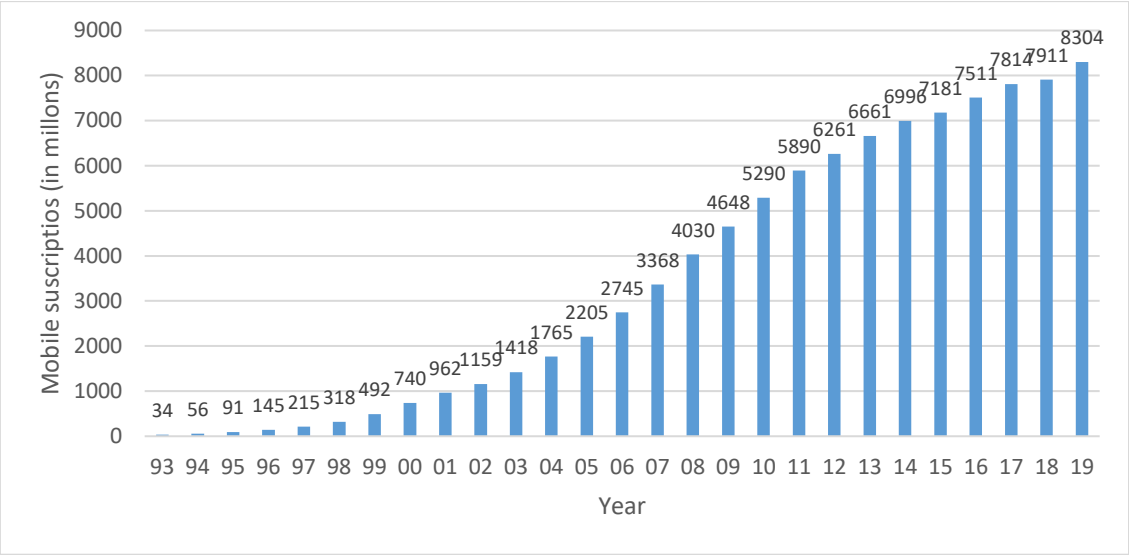
Figura 3. Valor de marca de las marcas más valiosas de telcomunicaciones a nivel mundial en 2019



Fuente: Statista.

El número de suscripciones a servicios de telefonía móvil ha ido aumentando notablemente con el pasar de los años, como indica la Figura 4. Pero en los últimos años se observa que la tasa de crecimiento ha ido desacelerando:

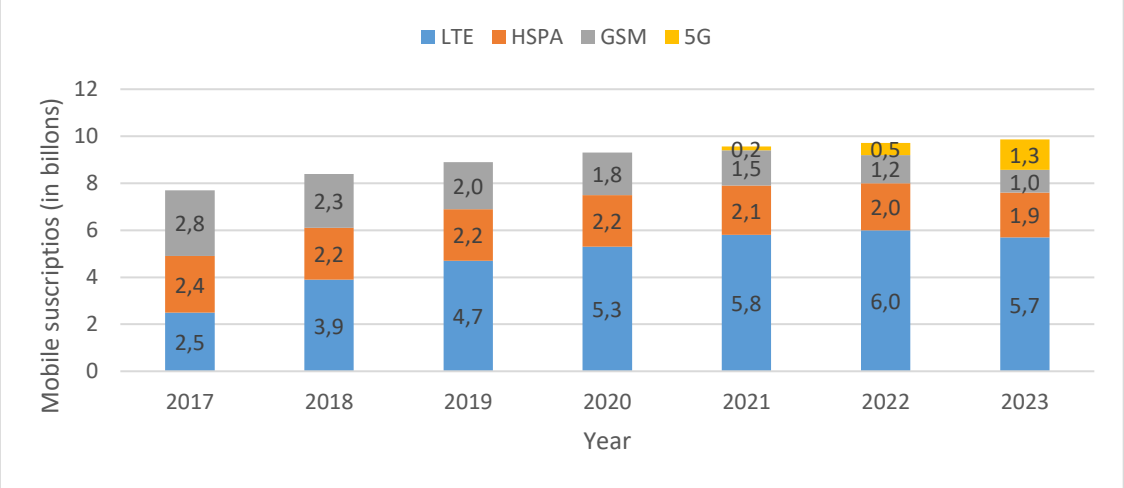
Figura 4. Número de suscripciones móviles a nivel mundial desde 1993 a 2019



Fuente: Statista.

En cuanto al tipo de tecnología que se utiliza en esta industria, se espera que para 2023 existan 1,3 billones de suscripciones al servicio de tecnología 5G, como exhibe la Figura 5 a continuación:

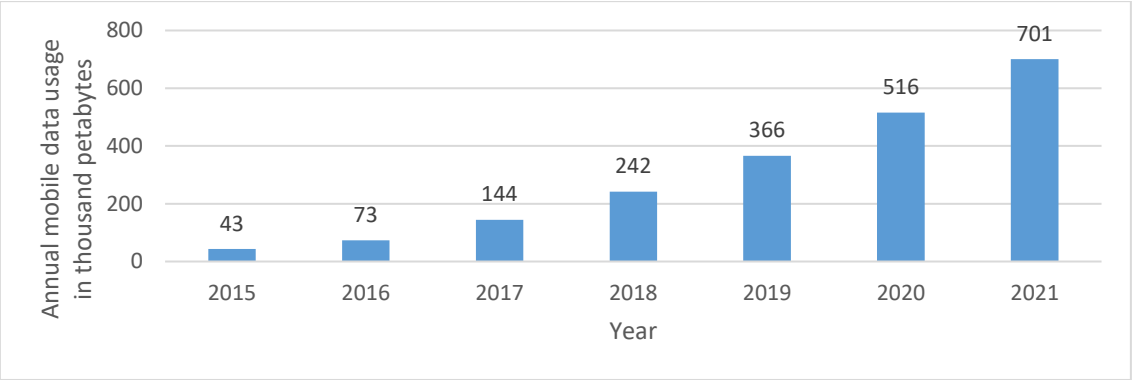
Figura 5. Número de suscripciones móviles a nivel mundial por tecnología desde 2017 a 2023



Fuente: Statista.

El uso de datos (servicio de internet en el dispositivo) ha crecido a gran escala, tanto por la provisión del servicio a una velocidad mayor en comparación con años anteriores y el número de dispositivos por los cuales se puede acceder a este servicio. Algunos valores del uso de datos se muestran en el siguiente gráfico de la Figura 6, con la aclaración que los valores de 2020 y 2021 son estimados:

Figura 6. Utilización de datos móviles a nivel mundial desde 2015 a 2021 (en miles de petabytes)

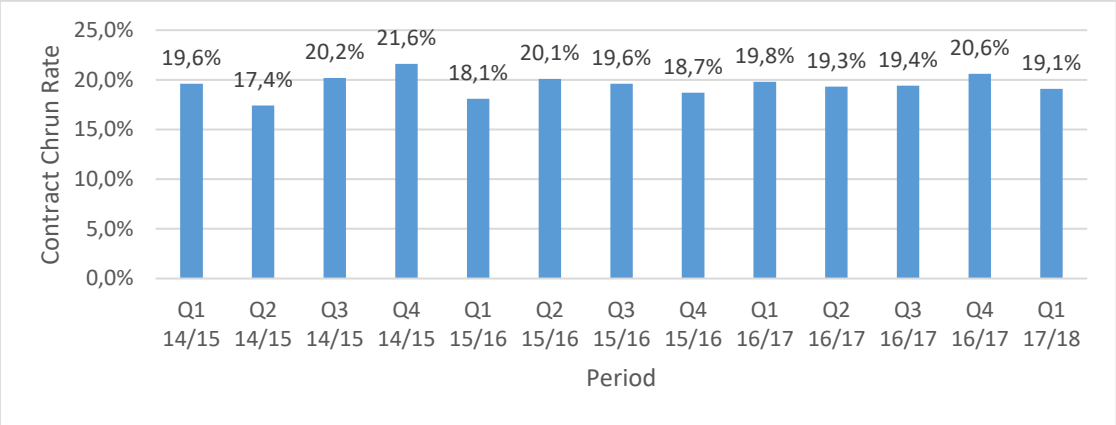


Fuente: Statista.

1.3.1.2 España y Europa

En lo que respecta al mercado español se puede visualizar en la Figura 7 la variación que mantiene la tasa de abandono para una compañía de telecomunicaciones (Vodafone) en lo que es el periodo 2014 a principio de 2018:

Figura 7. Tasa de abandono de Vodafone (contrato) en España



Fuente: Statista.

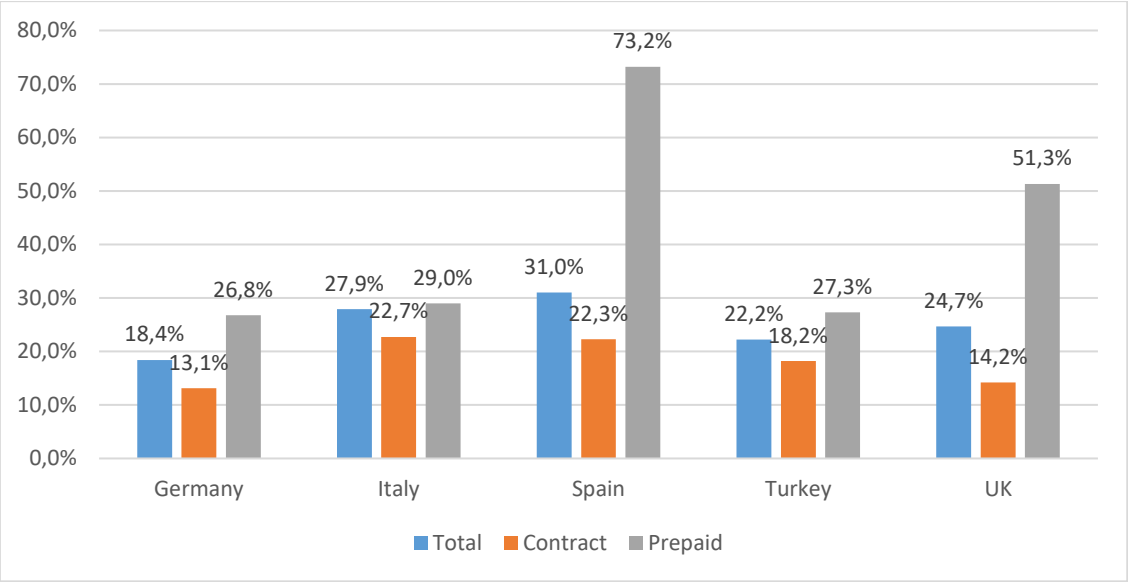
Como puede verse, existe un promedio cercano al 20% de "churn rate" por trimestre para esta compañía.

Otro dato es que, en octubre de 2019, casi 1 millón de personas realizaron el trámite de portabilidad, con el objetivo de cambiar de compañía de servicios móviles[1]. Esto demuestra que en la actualidad sustituir un operador de red por otro es un trámite sencillo y que, en principio, no conlleva una exorbitante suma de dinero en cuanto a gastos de transferencia. Además, evidencia que la oferta es amplia y las personas pueden escoger "libremente" a su operador de red en forma rápida.

Con respecto a Vodafone, pero a nivel europeo, los valores de "churn rate" que presenta para el cuarto trimestre de 2019 son:

Figura 8. Tasa de abandono de Vodafone en segmentos de comunicación móvil en el cuarto trimestre de 2019

[1] https://www.adslzone.net/2020/01/31/record-portabilidad-movil-octubre-2019/



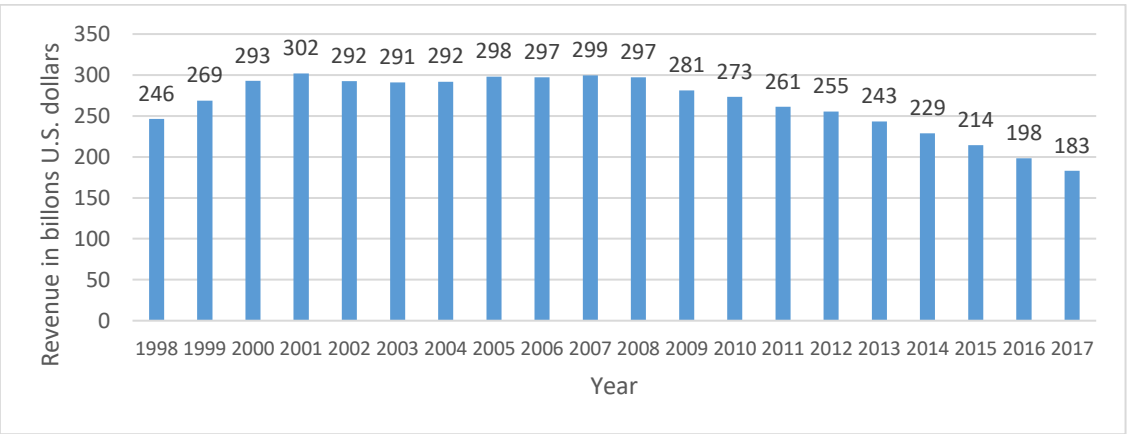
Fuente: Statista.

Es evidente que el mercado de prepago se caracteriza por una tasa de "churn" más alta que el mercado contractual, como se observa en la Figura 8. La causa de ello puede deberse a las diferentes barreras de salida que tiene una forma de contratación del servicio y la otra, pudiendo tener costes de penalización por abandono temprano en el caso del servicio por contrato.

1.3.1.3 Estados Unidos

El mercado de telecomunicaciones en Estados Unidos ha generado un promedio de 266 billones de dólares en los últimos años y es uno de los más grandes a nivel mundial, como puede detectarse en la Figura 9.

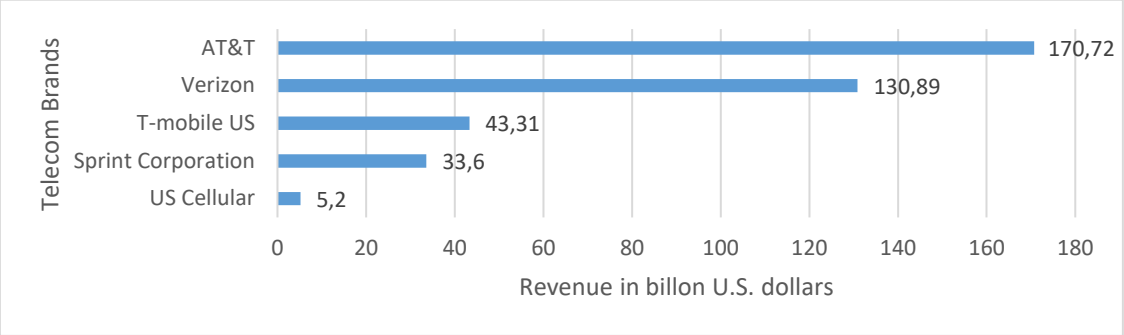
Figura 9. Ingresos en mil millones de dólares



Fuente: Statista.

La marca que mayores ingresos generó en 2018 en Estados Unidos fue AT&T con un total de 170,2 billones de dólares. Se exhiben estos valores comparados con otras compañías de Estados Unidos en la Figura 10.

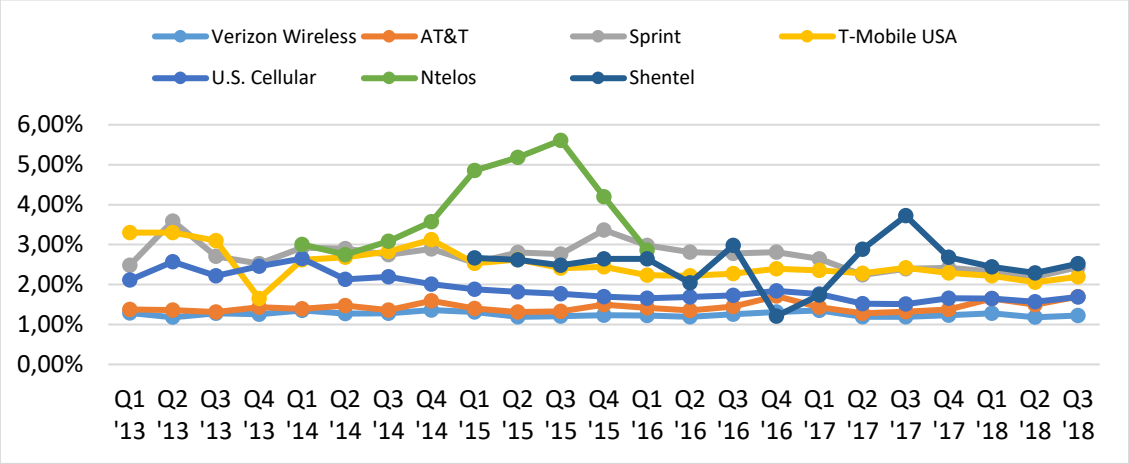
Figura 10. Ingresos de las mayores compañías de telecomunicaciones de EE.UU. en 2018



Fuente: Statista.

En cuanto al "churn rate" de las compañías de Estados Unidos se puede observar la variación que ha tenido la misma en los últimos años en la Figura 11.

Figura 11. Promedio mensual de tasa de abandono de proveedores de servicios móviles en EE.UU.



Fuente: Statista.

El promedio de tasa de abandono de la industria, o una empresa promedio de esta, se encuentra alrededor del 2,15% por trimestre.

El promedio acumulado anual responde a un valor cercano al 13%, promediando los años que se visualizan en el gráfico anterior.

En 2018 el número de suscripciones a operadores de red ascendió a 450 millones. Ese mismo año el mercado alcanzó los 182,78 mil millones de dólares en ingresos. Esto significa que, en promedio, cada suscripción aportó 406,17 dólares en ingresos en dicho año.

Si la tasa de abandono para dicho año se encontraba alrededor del 12%, los ingresos que perdieron las compañías debido a este fenómeno ascendieron a casi 22 mil millones de dólares para dicho año. De todas formas, esto no significa que la industria perdió esa cantidad de dinero, sino que dichos clientes cambiaron de proveedor. Pero esta perdida monetaria ha implicado que las empresas han tenido que invertir aún más para poder

captar nuevos clientes, cuestión que la bibliografía resalta como una acción que implica utilizar aún más recursos financieros.

Aun así, estos valores muestran la importancia que tiene la retención de clientes, o en contraposición, la relevancia que toma la tasa de abandono.

2 Estado del arte

El estudio del abandono de clientes y de la búsqueda de modelos que permitan predecir su probabilidad no constituye una novedad en el campo de la ciencia.

Muchos autores y autoras han desarrollado modelos, con distintos enfoques y diferentes formas de abordaje, para investigar este fenómeno que, con el pasar de los años, parece tomar una mayor popularidad.

Se ha estudiado el "churn rate" en diferentes sectores:

- Banca (Carmona et al., 2019);
- Seguros (Bolancé et al., 2016);
- Servicios financieros (Glady et al., 2009);
- Telecomunicaciones (Amin, Shah, et al., 2019; Aydin & Özer, 2005; Hassouna et al., 2015; Huang et al., 2012; Kim & Yoon, 2004; Núñez, 2015; Verbeke et al., 2012).

Dichas investigaciones han utilizado:

- Diferentes combinaciones de tratamientos de datos y su efecto en la predicción (Coussement et al., 2017; Neslin et al., 2006)
- Algoritmos predictivos (Amin et al., 2017; Hassouna et al., 2015; Lemmens & Croux, 2006; Tamaddoni et al., 2016; Tamaddoni Jahromi et al., 2014)
- Cantidades y variedades de conjuntos de datos (Amin, Shah, et al., 2019)
- Formas de comparaciones (Devriendt et al., 2019; Glady et al., 2009; Shaaban et al., 2012; Vafeiadis et al., 2015)
- Formas de capitalización de sus hallazgos (Ascarza, s. f.).

Algunos han utilizado métodos basados en árboles (Vafeiadis et al., 2015; Verbeke et al., 2012) ; otros métodos de ensamblados (Carmona et al., 2019; Lemmens & Croux, 2006); modelos de actualización en tiempo real (Balle et al., s. f.).

Aunque algunos de estos estudios han tomado una perspectiva de marketing (Jones & Sasser, s. f.; Devriendt et al., 2019) y han buscado establecer un criterio de uso empresarial de las probabilidades de abandono de cada cliente, no se han encontrado investigaciones que utilicen la segmentación estratégica del área de marketing como complemento.

En cuanto a los datos que utilizan, solo se ha encontrado una publicación que utiliza, entre otros, el mismo conjunto de datos que se utiliza en este trabajo. Los autores del trabajo son Adnan amin, Feras Al-Onbeidat, Babar Shah, Awais Adnan, Jonathan Loo y Sajid Anwar (Amin, Al-Obeidat, et al., 2019).

En dicha publicación presentan una aproximación al problema de la predicción de abandono de clientes basada en el concepto de la certeza en la estimación del modelo clasificador utilizando factores de distancia. También utilizan el clasificador de "Naive Bayes" como algoritmo base de comparación contra los que proponen.

A pesar de utilizar los mismos datos, en dicho trabajo no se han utilizado todas las variables que existen para este conjunto de datos, ya que solo han utilizado 21 variables.

El trabajo realiza un estudio con otra perspectiva al del presente trabajo de fin de máster, con métodos diferentes y conceptualizaciones de otro tipo para el abordaje del "Customer Churn Prediction (CCP)".

Por otro lado, existen trabajos de diferentes personas en la plataforma Kaggle, que han sido realizados sobre el mismo conjunto de datos del presente trabajo.

El caso del trabajo más desarrollado[2] sobre un conjunto de datos similar, pero con menor número de variables, se puede resumir en: exploración de las variables, procesamiento de las mismas y en la generación de diferentes modelos con el uso de varios algoritmos de clasificación.

Las principales diferencias entre dicho trabajo y este trabajo de fin de master son:

- En este TFM (trabajo de fin de master) se realizan modelos ensamblados sobre los obtenidos en forma aislada por diferentes algoritmos.
- También en el presente trabajo se utiliza un mayor número de variables iniciales, por lo que el conjunto de datos es más complejo que el utilizado en el trabajo de Kaggle.
- En el trabajo en Kaggle mencionado se utiliza una técnica de balanceo de las clases de la variable objetivo, específicamente la técnica SMOTE ("Synthetic minority oversampling technique").
- En dicho trabajo se utiliza una separación del conjunto de datos en entrenamiento y prueba. En este TFM se utiliza validación cruzada repetida para evitar el sobreajuste.
- En este TFM se obtienen mejores resultados en base a la predicción del abandono de cliente sin la utilización de técnicas de balanceo de muestra.
- En este TFM se utilizan las predicciones obtenidas para realizar acciones comerciales de retención de clientes, brindando una perspectiva de negocios más amplia al problema en cuestión.

[2] https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction

3 Objetivos

3.1 Objetivo Principal

Desarrollar modelos de predicción de probabilidad de abandono de los clientes de una compañía de telecomunicaciones, utilizando diferentes variables de tipo demográficas, de uso de los servicios y vinculación con la compañía.

3.2 Objetivos Secundarios

- Segmentar el conjunto de clientes en clusters que sean heterogéneos entre sí.
- Proponer acciones comerciales orientadas a mantener aquellos clientes que, bajo una estrategia corporativa, se ajusten al deseo de cartera de clientes de la organización.
- Detectar las variables que mayor poder de predicción obtengan, comparando los diferentes algoritmos utilizados.

4 Datos y Metodología

4.1 Conjunto de datos

Debido a que el interés de este trabajo es encontrar el mejor modelo de predicción de abandono de cliente y determinar cómo priorizar las acciones de retención sobre los más propensos a desertar, se utilizan datos[3] correspondientes al sistema CRM (Customer Relationship Management system) de una compañía que provee servicios de telecomunicaciones, TV e internet.

Los datos corresponden a una empresa de telecomunicaciones y servicios asociados, la cual provee a sus clientes de servicios de comunicación móvil, internet, televisión de cable, servicio de internet en el hogar, servicios de streaming, entre otros.
Este conjunto de datos corresponde a clientes que residen en diferentes ciudades del estado de California, Estados Unidos.

Para este estudio no es importante el nombre de los clientes, por lo que se los identifica a los mismos a través de un código de identificación (Customer_ID).

4.2 Metodología SEMMA

El instituto de SAS define a la minería de datos como “el proceso de muestrear (Sampling), explorar (Exploring), modificar (Modifying), modelar (Modeling) y valorar (Assessing) una gran cantidad de datos para descubrir patrones desconocidos que pueden ser utilizados como una ventaja competitiva”[4].

La primera letra de las 5 etapas que se han comentado antes conforman el acrónimo SEMMA, nombre que se utiliza para dicha metodología.
Realizando un acercamiento a cada una de estas etapas podemos comentar lo siguiente:

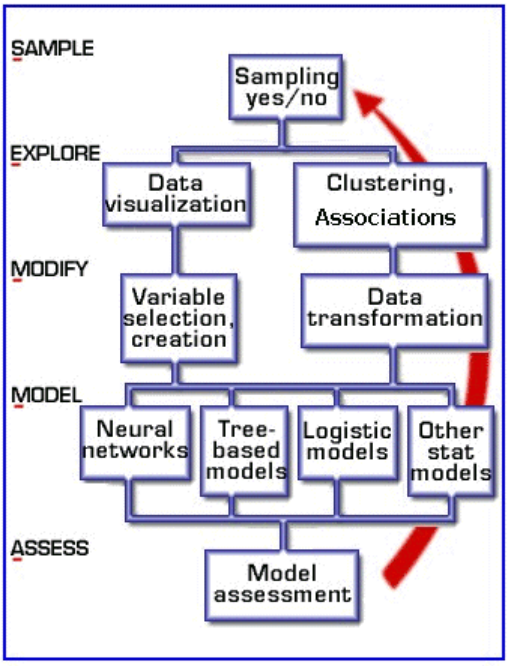
- Sampling (muestreo): realizar una muestra representativa del problema que se desea estudiar y utilizar particiones de entrenamiento-validación-testeo, en el

[3] Datos obtenidos de Kaggle: https://www.kaggle.com/ylchang/telco-customer-churn-1113#Telco_customer_churn_status.xlsx
[4] Frase traducida por el autor obtenida de https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

caso que sean necesarias cada una de estas partes, para efectuar un proceso correcto de ejecución de modelos y evitar tomar decisiones sobre modelos sobreajustados o subajustados (overfitting y underfitting respectivamente).

- Exploring (exploración): búsqueda de relaciones entre las diferentes variables, detección de datos anómalos (atípicos), estudio de las distribuciones de las variables, correlaciones entre las mismas, verificación de representatividad de las diferentes categorías, entre otros.
- Modifying (Modificación): aplicación de acciones que permitan realizar una "limpieza" de los datos, realizando diferentes tratamientos a los datos atípicos (outliers) y datos ausentes (missing values), así como la transformación de las variables, si correspondiera, en pos de encontrar la mejor relación de estas con la variable objetivo del problema en cuestión.
- Modeling (Modelización): desarrollo de modelos de predicción de la variable de interés (objetivo), utilizando diferentes técnicas de machine learning y variaciones de estas con sus respectivas parametrizaciones
- Assessing (Valoración / Evaluación): Comparación de la calidad de predicción de los modelos generados y pruebas individuales a los mismos. Se evalúan los resultados obtenidos en cuanto a su utilidad y consistencia como sección final del proceso de minería de datos.

Figura 12. Metodología SEMMA según el instituto SAS



Fuente: SAS.

La Figura 12 resume el proceso que implica la metodología SEMMA.

En lo que respecta a la etapa 4, modelado, se pueden utilizar diferentes técnicas de machine learning según el problema que se desee estudiar. En este caso, como la variable objetivo es de tipo binaria, es decir, que se estudiará la probabilidad de que un

cliente permanezca o no permanezca en la compañía, se utilizarán técnicas que permitan generar un resultado binario.

En las siguientes secciones se presentan los modelos a utilizan para el estudio del "churn" de los clientes.

4.3 Técnicas de modelado

4.3.1 Regresión Logística

El modelo de regresión logística busca obtener la probabilidad de las diferentes categorías de una variable basándose en funciones lineales de las variables dependientes. Al ser probabilidades, el objetivo que persigue es que las mismas se encuentren entre los valores 0 y 1.

En caso que la variable sea binaria, se plantea una variable objetivo de tipo "dummy", la cual expresa la aparición o no del evento bajo análisis. En el caso de este trabajo, la variable "Churn" representa el abandono (valor 1) o no abandono (valor 0) de un cliente.

El modelo de regresión logística buscará obtener la probabilidad de ocurrencia del evento de interés, es decir, la probabilidad que tiene cada uno de los clientes de abandonar la compañía, en función de las variables independientes con las que se cuenta.

Probabilidad de que suceda el evento (Y=1):

$$P(Y = 1|x_1, x_2, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

Un concepto relacionado es el de *odds* y este se explica como la probabilidad de ocurrencia del evento de interés en contraposición la probabilidad de que dicho evento no suceda:

$$odds(Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

Existe el concepto de *odds-ratio* que se define "como el cociente entre los *odds* de un suceso bajo una determinada confición y el *odds* de ese mismo suceso bajo otra condición, lo que permitirá evaluar el efecto de dichas condiciones sobre las probabilidades del suceso" (Calviño, 2019).

4.3.2 Redes neuronales

Las redes neuronales artificiales son modelos no lineales que se inspiran en el comportamiento de las neuronas humanas. Funcionan a través de diferentes tipos de capas y nodos conectados que buscan obtener un resultado determinado.

En cada una de estas capas se encuentran nodos con diferentes tipos de datos. Existen los nodos input, que son las variables que introducimos en el modelo; los nodos de la capa "oculta" o "hidden", que son los encargados de utilizar en diferentes ponderaciones la información de los nodos input, a través de pesos, y transformarlos mediante diferentes funciones; y la capa de nodos output que transforma los resultados de los nodos anteriores en un resultado pertinente al problema que se esté deseando resolver.

Los parámetros que pueden utilizarse para controlar el comportamiento de las diferentes redes son:

- Número de nodos
- Número de iteraciones de la red
- Tasa de aprendizaje (learning rate)
- Función de activación

4.3.3 Arboles de decisión

Los árboles de decisión, tanto de clasificación como de regresión, permiten segmentar los datos en función del uso de reglas simples, que se utilizan en forma secuencial a partir de las variables independientes del conjunto de datos.

Se inicia con un nodo raíz, que contiene todos los datos, y se realizan divisiones de los mismos utilizando diferentes criterios con las variables que se cuentan. Cada división de los datos corresponde a una submuestra del conjunto original, la cual se guarda momentáneamente en "nodos". La importancia de este método radica en obtener subconjuntos de datos homogéneos con respecto a la variable objetivo y heterogéneos con respecto a los demás nodos. Cada división de nodo se realiza bajo una segmentación "óptima", que busca utilizar un punto de corte que permita minimizar el error cometido en predecir la variable objetivo.

4.3.4 Bagging

Con el propósito de mejorar la capacidad predictiva de los árboles de decisión, surge la idea de combinar las predicciones de muchos árboles generados.

"Bagging" es un método basado en árboles, que utiliza técnicas de remuestreo con o sin reemplazamiento (Bootstrap) y genera, con estos distintos subconjuntos de datos, diferentes árboles de decisión. Una vez que se ha creado un número determinado de árboles, se promedian las predicciones obtenidas en cada uno de ellos.

Parámetros a controlar en bagging:

- Tamaño de muestra
- Si se utiliza reemplazamiento (Bootstrapping) o no
- Número de iteraciones (árboles) a promediar
- Características de los árboles:
    - Número de hojas finales o profundidad del árbol
    - Número de divisiones máxima en cada nodo
    - Número de observaciones mínimo en un nodo

4.3.5 Random Forest

Es una modificación al bagging. Lo interesante de este método es que agrega la posibilidad de variar el uso de las variables en cada árbol a generar.

Permite utilizar un número menor del total de variables disponibles para generar cada uno de los árboles. La ventaja de esta incorporación es que permite ampliar la capacidad de generalización de los datos, evitando que ciertas variables tomen un papel

protagonista por sobre las demás y que se deba realizar una selección rígida previa de las mismas.

Los parámetros a controlar en random forest son:

- Tamaño de muestra
- Si se utiliza reemplazamiento o no
- Número de iteraciones (árboles) a promediar
- Número de variables a muestrear en cada nodo
- Características de los árboles:
    - Número de hojas finales o profundidad del árbol
    - Número de divisiones máxima en cada nodo
    - Número de observaciones mínimo en un nodo

4.3.6 Gradient Boosting

Este método permite ir actualizando las predicciones en la dirección de decrecimiento dada por el negativo del gradiente.

Es decir, que utilizando árboles de decisión, ajusta las predicciones en base al error generado por un árbol. Con el residuo anterior, realiza una nueva predicción en dirección al signo del error.

Es un proceso iterativo, que permite ajustar los diferentes arboles generados a los datos que se utilizan para la construcción del modelo.

Los parámetros a utilizar en gradient boosting son:

- Constante de regularización (shrink)
- Número de iteraciones
- Características de los árboles:
    - Número de hojas finales o profundidad del árbol
    - Número de divisiones máxima en cada nodo
    - Número de observaciones mínimo en un nodo

4.3.7 Extreme Gradient Boosting

Es una ampliación del algoritmo de gradient boosting, al cual se le adiciona el uso de la regularización, que es una técnica orientada a la reducción de la varianza de los errores. Esta regularización se utiliza en la optimización interna del algoritmo, durante el proceso de estimación de los parámetros. Explicado de otra forma, se introduce un término de penalización por el uso de parámetros con valores más altos, como el uso un número alto de hojas y el resultado de la predicción de cada una de estas.

Parámetros a controlar:

- Los mismos que en gradient boosting
- Regularización:
    - Alpha
    - Lambda

4.3.8 Support Vector Machine

Las Máquinas de Soporte Vectorial o Support Vector Machines se esmeran en separar en forma lineal las clases de la variable objetivo, utilizando métodos algebraicos que buscan el hiperplano de separación.

Debido a que no es usual que la separación perfecta de los datos sea descubierta o exista, se utiliza lo que se denomina como “soft margin”, es decir, se permite un porcentaje de observaciones mal clasificadas por los separadores para evitar un sobreajuste del modelo a los datos de entrenamiento.

Puede ocurrir que la separación de las clases no se corresponda con un problema lineal, y es por ello que se utiliza el “truco kernel”. Este concepto permite “trabajar en un espacio de dimensión superior donde si tenga sentido la separación lineal. Simplemente extrapolar los datos con más dimensiones nos permite encontrar separadores lineales” (Portela, 2020).

La función kernel permite transformar el espacio en el que se encuentran los datos.

Por ejemplo, cuando el kernel es lineal permite realizar una división de los datos con una función lineal. Existe otro caso, el uso del kernel polinómico, que permite realizar distinciones en espacios no lineales, a través del uso del parámetro de grado de polinomio (*degree*).

El otro caso conocido de SVM es el kernel RBF (Radial Basis Function), que se suele utilizar en problemas de clasificación binaria, debido a que suele ser más flexible que el kernel polinomial.

Parámetros a considerar:

- Parámetro “C”
- Grado del polinomio y la escala (en SVM polinomial)
- Sigma (en el caso de SVM RBF)

4.3.9 Ensamblado

Un modelo ensamblado es un promedio de otros modelos obtenidos anteriormente, que busca reducir el error en la predicción de la variable objetivo. En lugar de obtener un único modelo con una gran precisión, se generan modelos ensamblados mediante la combinación de diferentes modelos, con el propósito de obtener una sinergia positiva en dicha mezcla de algoritmos.

Algunos de los algoritmos anteriormente definidos son métodos de ensamblados en sí mismos, como bagging, random forest y gradient boosting. Esto se debe a que combinan un gran número de árboles de decisión y realizan una predicción con un promedio de estos.

4.3.10 Validación cruzada repetida

Se utilizará validación cruzada repetida con las técnicas de machine learning mencionadas anteriormente. Este proceso de validación consiste dividir los datos en *k* grupos, dejando uno de estos grupos en forma separada del resto, que no será utilizado

para la construcción del modelo. El resto de los grupos de datos son los que efectivamente se utilizan para crear el modelo en cuestión. Este proceso descrito se realiza *n* veces y del mismo se extrae el error de predicción utilizando el grupo aislado previamente como conjunto de prueba.
Para este trabajo se han utilizado 10 grupos y 10 repeticiones.

4.3.11 Medidas de comparación
Los modelos generados serán comparados entre si bajo diferentes medias, como es la exactitud de la precisión (Accuracy), el área bajo la curva ROC y la tasa de fallos en la predicción (misclassification rate). Con el proceso de validación cruzada repetida y las medidas mencionadas, los resultados serán facilitados a través de gráficos de caja (boxplot) con el objetivo de cotejar los mismos en forma visual.

4.4 K-means
4.4.1 Definición
K-means, o k-medias, es un algoritmo no supervisado, que busca agrupar un conjunto de datos de *n* observaciones en *k* grupos en el que cada observación pertenece al grupo cuya media sea más cercana.

Lo interesante de este algoritmo es que uno es quien escoge el número de grupos (*k*) y k-means, mediante un proceso iterativo, busca los mejores centroides de los grupos para poder dividir los datos en esos *k* grupos determinados.

Este algoritmo será utilizado para agrupar o segmentar a los clientes del conjunto de datos y será un complemento al mejor modelo escogido bajo los algoritmos supervisados explicados anteriormente.

5 Descripción de los datos

5.1 Definición

Los datos, como ya se ha comentado, pertenecen a una muestra extraída del sistema CRM de una compañía de telecomunicaciones, y servicios asociados, de Estados Unidos.

El archivo en el cual se encuentra esta información, en formato xlsx, contiene un total de 7043 filas, cada una correspondiente a un cliente de la compañía.

La variable objetivo ("Churn") es de tipo binaria, la cual determina si, al final del trimestre, el cliente ha abandonado o no la compañía. En el caso que la misma obtenga el valor 0 significa que el cliente al terminar dicho trimestre aún mantiene su relación comercial con la compañía; en caso que dicho valor sea 1 indica que la persona ha abandonado a la empresa.

Las variables con las que se cuenta difieren tanto en su formato como en su "naturaleza". Son variables de tipo demográficas, de ubicación, servicios contratados, consumo y uso, y el estado de la relación contractual entre el cliente y la empresa.
A modo de resumen los tipos de variables se dividen en:

- 21 binarias
- 14 categóricas
- 14 numéricas
- 3 de geolocalización
- 3 unarias
- 1 identificadora

A continuación, se exhiben todas las variables en la Tabla 1, con su interpretación correspondiente:

Tabla 1. Descripción de Variables

Variable	Grupo de variable	Tipo	Explicación
CustomerID	ID	Identificador	Identificador de cada cliente
Partner	Demográfica	Binaria	Indica si el cliente vive con otra persona
Gender	Demográfica	Binaria	Género del cliente: Masculino o Femenino
Age	Demográfica	Numérica	Edad del cliente
Under 30	Demográfica	Binaria	Indica si el cliente es una persona menor a 30 años
Senior Citizen	Demográfica	Binaria	Indica si el cliente es una persona mayor a 65 años
Married	Demográfica	Binaria	Indica si la persona se encuentra casada
Dependents	Demográfica	Binaria	Indica si la persona tiene alguna persona a su cargo que vive con ella
Number of Dependents	Demográfica	Categórica	Indica la cantidad de personas dependientes que viven con el cliente
Device Protection	Servicio	Categórica	Indica el tipo de servicio de protección de dispositivos que el cliente ha contratado
Tech Support	Servicio	Categórica	Indica si el cliente ha contratado el soporte técnico
Monthly Charges	Servicio	Continua	Cargos mensuales por servicios contratados
Referred a Friend	Servicio	Binaria	Indica si el cliente ha referido a otra persona para que sea cliente
Number of Referrals	Servicio	Categórica	Indica el número de personas que ha referido el cliente

Tenure in Months	Servicio	Continua	Número de meses que el cliente ha permanecido en la compañía
Offer	Servicio	Categórica	Indica la última oferta que el cliente ha aceptado
Phone Service	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de teléfono en su hogar
Avg Monthly Long Distance Charges	Servicio	Continua	Cargos promedio por la realización de llamadas de larga distancia
Multiple Lines	Servicio	Binaria	Indica si el cliente tiene más de una línea de teléfono contratada
Internet Service	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de internet
Internet Type	Servicio	Categórica	Indica qué tipo de conexión de internet tiene contratado
Avg Monthly GB Download	Servicio	Continua	Cantidad promedio de Gigabytes utilizados o descargados
Online Security	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de seguridad online
Online Backup	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de respaldo online
Device Protection Plan	Servicio	Binaria	Indica el tipo de protección de dispositivos que tiene contratado el cliente
Premium Tech Support	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de soporte técnico adicional
Streaming TV	Servicio	Binaria	Indica si el cliente utiliza servicios de streaming para ver programas de TV
Streaming Movies	Servicio	binaria	Indica si el cliente utiliza servicios de streaming para ver películas
Streaming Music	Servicio	Binaria	Indica si el cliente utiliza servicios de streaming para escuchar música
Unlimited Data	Servicio	Binaria	Indica si el cliente tiene contratado el servicio de datos ilimitados
Contract	Servicio	Categórica	Tipo de contrato actual
Paperless Billing	Servicio	Binaria	Indica si el cliente utiliza factura electrónica
Payment Method	Servicio	Categórica	Modo de pago
Monthly Charge	Servicio	Numérica	Cargos mensuales
Total Charges	Servicio	Numérica	Cargos totales
Total Refunds	Servicio	Numérica	Reembolsos
Total Extra Data Charges	Servicio	Numérica	Cargos extra por uso de datos por fuera del plan contratado
Total Long Distance Charges	Servicio	Numérica	Cargos extras por llamadas de larga distancia
Total Revenue	Servicio	Numérica	Sumatoria de todos los cargos
Quarter	Servicio	Unaria	Trimestre al que corresponde la información
Country	Ubicación	Unaria	País de residencia del cliente
State	Ubicación	Unaria	Estado de residencia del cliente
City	Ubicación	Categórica	Ciudad de residencia
Zip Code	Ubicación	Categórica	Código ZIP de la ciudad de residencia
Lat Long	Ubicación	Geo	Ubicación geográfica
Latitude	Ubicación	Geo	Latitud
Longitude	Ubicación	Geo	Longitud
Population	Ubicación	Numérica	Población de la ciudad de residencia del cliente
Satisfaction Score	Estado	Categórica	Indicador de satisfacción con la compañía provista por el cliente
Customer Status	Estado	Categórica	Indica el estado del cliente al finalizar el trimestre
Churn Label	Estado	Binaria (Objetivo)	Indica si el cliente ha abandonado o no la compañía al final del trimestre (“No” / “Yes”)
Churn Value	Estado	Binaria (Objetivo)	Indica si el cliente ha abandonado o no la compañía al final del trimestre (“0” / “1”)

Churn Score	Estado	Numérica	Valor de probabilidad de abandono (obtenido del CRM)
CLTV	Estado	Numérica	Valor del cliente para la compañía
Churn Category	Estado	Categórica	Categorías de motivos de abandono
Churn Reason	Estado	Categórica	Motivo de abandono de la compañía

5.2 Variable Objetivo

La variable objetivo ("Churn Value") cuenta con una proporción diferente de valores ceros y unos, como es de esperar en este tipo de casos.

De las 7043 observaciones que se disponen, 5174 de ellas tienen el valor cero en la variable objetivo, es decir, que corresponden a clientes que al finalizar el trimestre bajo análisis no han abandonado la compañía.

Por el otro lado, las 1869 observaciones restantes corresponden a clientes que si han abandonado la compañía.

Estos valores reflejan una proporción de 73,46% de valores en cero (no abandono) y de 26,54% de valores en uno (abandono).

5.2.1 Comentarios adicionales

En este tipo de problema binario, en el cual el evento objetivo no suele tener una gran representación en muestras representativas de la población de datos, es habitual plantear la posibilidad de hacer uso de técnicas de balanceo, como por ejemplo oversampling.

Diferentes autores y autoras han realizado sus estudios utilizando distintas técnicas de balanceo de eventos de la variable objetivo, pero por otro lado existen otras personas que no han utilizado ninguna de estas.

Aun así, dentro grupo de personas que han utilizado estas técnicas, no todas afirman que el uso de oversampling o alguna técnica similar mejora sustancialmente la capacidad de predicción de los diferentes algoritmos empleados.

El hecho de emplear técnicas de balanceo para aumentar el peso del evento menos representado, disminuir la magnitud que toma el evento mayoritario o incluir información sintética para favorecer a la clase minoritaria implica modificar el estado natural de la variable bajo estudio. De todas formas, en varios casos, se utiliza un conjunto de entrenamiento con una representación de 50%/50% para el evento y el no evento de la variable dependiente, mientras que dentro del conjunto de prueba se mantiene la proporción original.

A modo de resumen se describe lo realizado por algunos autores de estudios de predicción binaria y las conclusiones sobre el balanceo de la variable objetivo:

- Se utilizan arboles de decisión sensibles a costes, entre otros algoritmos, con el objetivo de incorporar el desbalance de clases, pero aun así los mejores modelos obtenidos son la regresión logística y gradient boosting sin balanceo de clases (Tamaddoni Jahromi et al., 2014).

- Se ha utilizado un dataset con 12,55% de clientes que abandonan y el modelo ganador fue construido con el algoritmo “Adacost”, que es un algoritmo basado en arboles con una función de ajuste de coste en base a las observaciones mal clasificadas (Glady et al., 2009)
- Utiliza un conjunto de datos de 100.000 observaciones con 1,8% de clientes que dejan la compañía. Se utiliza una técnica de oversampling sobre el conjunto de entrenamiento y el mejor modelo obtenido es una regresión logística (Neslin et al., 2006).
- Se arriba a la conclusión de que el uso de oversampling generalmente no mejora el rendimiento de los modelos. En dicho estudio se utilizan varios conjuntos de datos, con un promedio de 3,20% de clientes que abandona, a excepción de uno, cuyo porcentaje es de 14,14% para el evento a predecir (Verbeke et al., 2012).
- Se utiliza la técnica de undersampling sobre la categoría mayoritaria (“Non-churn”) para el conjunto de entrenamiento. La representación de clientes que abandonan la compañía es de 3,28% (Huang et al., 2012).
- Se emplean diferentes datasets, siendo el mayor valor para los clientes que abandonan de 14,5%, y en un promedio global la representación del evento es de 5,51%. Se utilizan diferentes técnicas de balanceo y algoritmos. Se concluye que los resultados varían según las combinaciones realizadas. En el caso de algoritmos como bagging y random forest no se observa una ganancia sustancial de los modelos construidos bajo el uso de técnicas de balanceo y los modelos ensamblados son los que mejores resultados obtienen (Zhu et al., 2017).
- Recomiendan el uso de un esquema de muestra balanceada debido a que utilizan un conjunto de datos con una proporción de “churners” del 1,8% (Lemmens & Croux, 2006).
- No se realiza ningún tipo de balanceo a los datos, que se caracteriza por tener 14000 observaciones y una representación del 20% de clientes que abandonan la empresa (Bolancé et al., 2016).
- Utilizan dos conjuntos de datos: en uno el porcentaje de clientes que abandonan es del 13,25% y en el otro es de 25,52%. No utilizan ningún tipo de técnica de balanceo de eventos para estos datos (Devriendt et al., 2019).

Se puede detectar con la lista anterior que se han empleado varias formas para tratar este tema.

Debido a que la clase minoritaria (clientes que abandonan) de la variable objetivo de los datos que se utilizan en este trabajo se encuentra representada por encima del 20% del total, no se utiliza ninguna técnica de balanceo. Esta decisión es una medida similar a otros autores y autoras con este tipo de representación del evento de interés (Bolancé et al., 2016; Devriendt et al., 2019; Verbeke et al., 2012).

5.3 Análisis Descriptivo

Con el uso del programa SAS Enterprise Miner, se obtienen los estadísticos descriptivos correspondientes a las variables numéricas para las 7043 observaciones (Tabla 2):

Tabla 2. Estadísticos variables numéricas

Variable	Ausente	Mín.	Máximo	Media	Mediana	D.E.	Asimetría	Curt.
Total_Refunds	0	0	49,79	1,96	0,00	7,90	4,33	18,35
Total_Extra_Data_Charges	0	0	150,00	6,86	0,00	25,10	4,09	16,46
Total_Long_Distance_Charges	0	0	3564,72	749,10	401,44	846,66	1,24	0,64
Avg_Monthly_GB_Download	0	0	85,00	20,52	17,00	20,42	1,22	0,88
Total_Charges	0	18,80	8684,80	2280,38	1394,55	2266,22	0,96	-0,23
Population	0	11,00	105285,00	22139,60	17554,00	21152,39	0,91	0,33
Total_Revenue	0	21,36	11979,34	3034,38	2108,64	2865,20	0,92	-0,20
Tenure_in_Months	0	1,00	72,00	32,39	29,00	24,54	0,24	-1,39
Avg_Monthly_Long_Distance_Charge	0	0	49,99	22,96	22,89	15,45	0,05	-1,25
Monthly_Charge	0	18,25	118,75	64,76	70,35	30,09	-0,22	-1,26
Churn_Score	0	5,00	96,00	58,51	61,00	21,17	-0,16	-1,09
Age	0	19,00	80,00	46,51	46,00	16,75	0,16	-1,00
CLTV	0	2003,00	6500,00	4400,30	4527,00	1183,06	-0,31	-0,93

Tabla 3. Estadísticos variables categóricas

Variable	Categorías	Ausentes
CustomerID	7043	0
Lat Long	1679	0
Zip Code	1626	0
Latitude	1626	0
Longitude	1625	0
City	1106	0
Churn Score	81	0
Churn Reason	21	0
Number of Referrals	12	0
Number of Dependents	10	0
Offer	6	0
Churn Category	6	0
Satisfaction Score	5	0
Internet Type	4	0
Device Protection	3	0
Tech Support	3	0
Contract	3	0
Payment Method	3	0
Customer Status	3	0
Partner	2	0
Gender	2	0
Under 30	2	0
Senior Citizen	2	0
Married	2	0

Dependents	2	0
Referred a Friend	2	0
Phone Service	2	0
Multiple Lines	2	0
Internet Service	2	0
Online Security	2	0
Online Backup	2	0
Device Protection Plan	2	0
Premium Tech Support	2	0
Streaming TV	2	0
Streaming Movies	2	0
Streaming Music	2	0
Unlimited Data	2	0
Paperless Billing	2	0
Churn Label	2	0
Churn Value	2	0
Country	1	0
State	1	0
Quarter	1	0

Observando la Tabla 3, se detecta que no existen valores ausentes para ninguna de las variables que se han presentado.
De todas formas, esto no significa que todas las variables serán utilizadas.

Dentro del conjunto de datos existen tres variables que solo contienen un valor, siendo las mismas: Country, State y Quarter. La primera solo toma el valor “United States”; la segunda el valor de “California”; y la última el valor de “Q3” indicando que es el tercer trimestre del año. Estás variables no tendrán ningún tipo de participación en el presente trabajo, debido a que no aportan ningún tipo de valor predictivo.

Cómo es sabido, la variable “CustomerID” representa en forma única a cada registro del conjunto de datos, es por ello que tiene una cantidad de “categorías” igual a cada registro.

Realizando un acercamiento a las variables con más número de categorías, se puede apreciar que existen 5 variables con un número demasiado alto de categorías para la cantidad de observaciones que se disponen. Estas variables son: City, Longitude, Zip Code, Latitude y Lat Long. Estas variables están relacionadas entre sí, ya que representan la ubicación de cada uno de los clientes que existen dentro de la muestra.

Analizando la frecuencia de las 20 ciudades más repetidas que se observan en la variable “City” se observa lo siguiente:

Tabla 4. Top 20 Ciudades

Ciudad	Frecuencia Absoluta	Frecuencia Relativa
Los Angeles	293	4,16%

San Diego	285	4,05%
San Jose	112	1,59%
Sacramento	108	1,53%
San Francisco	104	1,48%
Fresno	61	0,87%
Long Beach	60	0,85%
Oakland	52	0,74%
Escondido	51	0,72%
Stockton	44	0,62%
Fallbrook	43	0,61%
Glendale	40	0,57%
Bakersfield	39	0,55%
Temecula	38	0,54%
Berkeley	32	0,45%
Riverside	32	0,45%
Pasadena	30	0,43%
Whittier	30	0,43%
Anaheim	28	0,40%
Irvine	28	0,39%

Dentro de la Tabla 4, “Los Ángeles” resulta la categoría con mayor número de apariciones, pero dichas apariciones representan menos del 5% de los datos disponibles. Debido a estas características de la variable, la misma no será tenida en cuenta para la creación de modelos predictivos de la variable objetivo. Esta determinación aplica también a las otras cuatro variables asociadas a “City”, comentadas anteriormente.

5.4 Depuración de datos

Además de las variables quitadas del análisis explicadas en la sección previa, existen otras variables que no serán tenidas en cuenta, debido a que reflejan la misma información que otras existentes en el conjunto de datos.

Existen dos variables, de tipo binaria, que están directamente asociadas a la variable “Age”. Las mismas son: “Under30” y “Senior Citizen”. Ambas variables son construcciones originadas en la variable que exhibe la edad de los clientes, razón por la cual serán eliminadas del análisis.

Por otro lado, tanto para describir la información de número de personas dependientes del cliente y el número de personas a los que ha referido el mismo ocurre una situación similar. De ambos casos se cuenta con dos tipos de variables: una binaria y una categórica. La primera de ella describe el hecho de si el cliente tiene dependientes o no, y si ha referido a la compañía a otra persona o no. La variable categórica describe el

grupo de personas dependientes, y por otro lado, el grupo de personas a las que ha referido el cliente. Es por ello que solo se contemplarán las variables “Number of Dependents” y “Number of Referrals” dentro del análisis, ya que agregan mayor valor que las variables binarias de “Dependents” y “Referred a Friend”. Es decir, que utilizarán las categóricas y no las binarias de ese conjunto de variables.

Casos similares al anterior ocurren para las variables: “Tech Support” y “Tech Support Premium”; “Internet Service” e “Internet Service Type”; “Device Protection” y “Device Protection Plan”. En todos los casos, se utilizará la segunda variable, es decir, la categórica de más de 2 grupos.

Otra variable que no será tenida en cuenta es “Total Revenue”, debido a que es una variable calculada en base a otras variables existentes. La fórmula que explica a “Total Revenue” es:

$$\text{Total Revenue} = Total\ \ Charges - Total\ Refunds + Total\ Extra\ Data\ Charges + Total\ Long\ Distance\ Charges$$

Como último paso de eliminación de variables, serán quitadas del análisis aquellas variables directamente relacionadas con la variable objetivo (“Churn Value”). Las mismas son: “Satisfaction Score”, “Customer Status”, “Churn Label”, “Churn Score”, “Churn Category” y “Churn Reason”.

Excepto “Satisfaction Score”, el resto de las variables anteriores se explican directamente por su nombre y descripción.

El caso de “Satisfaction Score” y su eliminación se debe a que la misma se encuentra linealmente relacionada con la variable objetivo.

Tabla 5. Relación "Churn" con "Satisfaction Score"

Churn	Categoría Satisfaction_score	Observaciones	Frecuencia Relativa
Yes	1	922	13,09%
	2	518	7,35%
	3	429	6,09%
No	3	2236	31,75%
	4	1789	25,40%
	5	1149	16,31%

De hecho, observando la Tabla 5, los valores de “1” y “2” de “Satisfaction Score”, que significan “Muy poco satisfecho/a” y “Poco satisfecho/a” implican un abandono de ese cliente al final del trimestre. Por el otro lado, los valores “4” y “5”, cuyo significado es “Bastante satisfecho/a” y “Muy satisfecho/a” respectivamente, implican que el cliente seguirá dentro de la compañía al final del trimestre. El caso del último valor restante, el valor “3”, que indica “Satisfecho/a”, contiene clientes que han abandonado la compañía al finalizar el período y otros que aún se mantienen en ella. La proporción de los clientes que se quedan en la compañía es muy superior en la categoría de “Satisfecho/a” con respecto a lo que la abandonan.

Es por ello que esta variable se elimina, debido a que la satisfacción está íntimamente relacionada con el abandono o la permanencia del cliente en la compañía, cuyo efecto suele depender de la industria en la cual se esté realizando el estudio (Jones & Sasser, s. f.).

Además, desde un punto de vista del negocio, no siempre se cuenta con el estado de satisfacción expreso del cliente y, en otras ocasiones, es difícil de obtener. Incluso cuando su magnitud es conocida puede implicar que es demasiado tarde para realizar acciones de retención, es decir, que si su puntaje es bajo, el cliente ya ha tomado la decisión de abandonar la compañía.

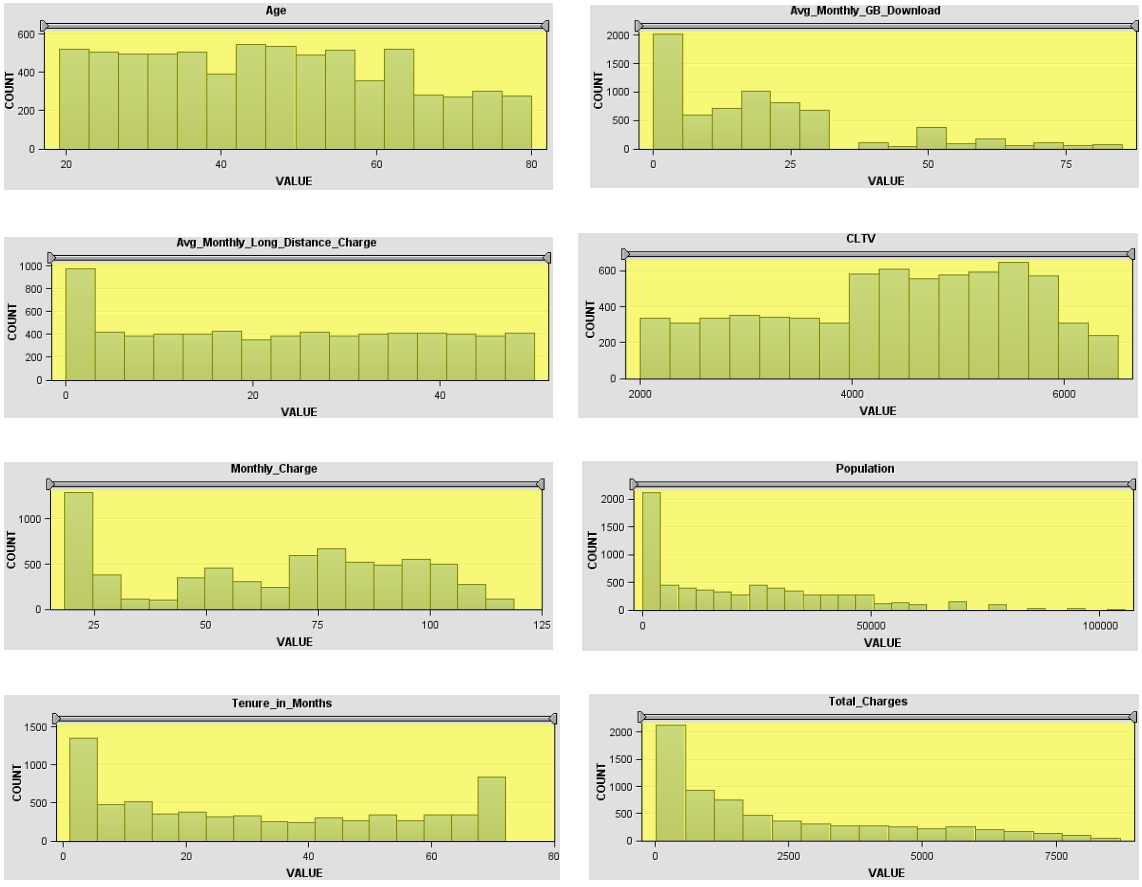
Debido a estas causas adicionales, no se utilizará esta variable debido a que el modelo obtenido será dependiente a ella y si no se cuenta con la misma para algún cliente el modelo obtenido puede no ser tan preciso como se desea.

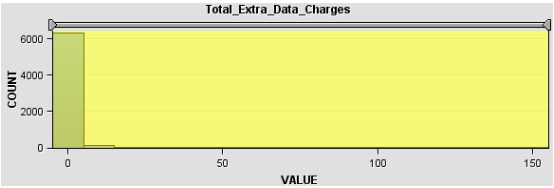
5.4.1 Valores atípicos

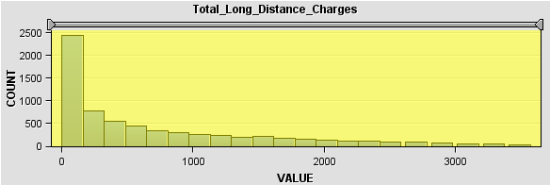
Se estudiará la distribución que presente cada una de estas variables con el objetivo de verificar si existen valores atípicos.

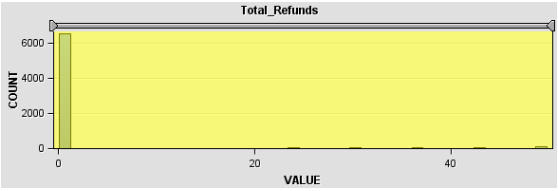
La distribución que presentan las variables se muestra en la Figura 13 a través de sus respectivos histogramas obtenidos de SAS Miner Enterprise:

Figura 13. Distribuciones de variables numéricas









Excepto la variable “Age” y “CLTV”, las demás presentan una distribución asimétrica. Para la detección de valores atípicos se utilizarán diferentes métodos límite, según cada una de las distribuciones anteriores que presenta cada variable.

Para el caso de las variables asimétricas que cuentan con su mediana igual a cero, se utilizará el método de percentiles extremos; para las que su mediana es diferente de cero se utiliza la desviación absoluta mediana (MAD en SAS Miner), tal como se indica en la Tabla 6.

Tabla 6. Variables numéricas y métodos límites utilizados

VARIABLE	MÉTODO LÍMITE UTILIZADO
AGE	Desviación estándar
AVG_MONTHLY_GB_DOWNLOAD	Desviación absoluta mediana (MAD)
AVG_MONTHLY_LONG_DISTANCE_CARGE	Desviación absoluta mediana (MAD)
CLTV	Desviación estándar
MONTHLY_CHARGE	Desviación absoluta mediana (MAD)
POPULATION	Desviación absoluta mediana (MAD)
TENURE_IN_MONTHS	Desviación absoluta mediana (MAD)
TOTAL_CHARGES	Desviación absoluta mediana (MAD)
TOTAL_EXTRA_DATA_CHARGES	Percentiles Extremos
TOTAL_LONG_DISTANCE_CHARGES	Desviación absoluta mediana (MAD)
TOTAL_REFUNDS	Percentiles Extremos

Luego de aplicar dichos métodos de límite, y convertir los valores atípicos detectados como valores ausentes, la única variable que ha presentado casos de atípicos es “Total_Refunds”, con 34 valores ausentes en total. Los valores que fueron convertidos son los que superaban el valor de 47,19 que corresponden a los que integran el último percentil de esta variable.

De esta forma, solo a esta variable se le ha aplicado en forma definitiva la transformación de los valores atípicos en ausentes.

Como paso siguiente, se han convertido estos valores ausentes utilizando la media de la variable, obteniendo un conjunto de datos limpios de valores ausentes y atípicos en lo que respecta a las variables numéricas.

También se han realizado pruebas de transformación de las variables continuación, con el objetivo de detectar alguna relación no visible entre estas y la variable objetivo. Se ha utilizado el nodo de Transformar variables de SAS Miner Enterprise, pero no se ha detectado una opción de transformación útil, por lo que las variables continuarán como han sido cargadas.

5.4.2 Agrupación de categorías

En cuanto a las variables de tipo binarias todas han sido reestructuradas. Los valores "No" han sido reemplazados por ceros y los valores "Yes" han sido reemplazados con unos, con el objetivo de simplificar su tratamiento posterior dentro de la creación de modelos predictivos de la variable objetivo.

La variable "Gender" ha sido convertida a "Gender_Female", indicando con el valor cero a clientes hombres y con el valor uno a clientes mujeres.

Por otro lado, se ha realizado un análisis de la representación que tiene cada categoría dentro del conjunto de datos.

Se ha tomado como valor mínimo tolerante la representación del al menos 5% de la categoría dentro de los datos que se disponen. Es decir, que si la categoría no alcanza el valor de 5% será reagrupada.

Para las variables "Contract", "Internet_Type", "Offer" y "Payment Method" no se han realizado re-agrupaciones de variables.

La variable "Number of Dependants" si ha sufrido una re-estructuración, debido a que las categorías que corresponden a los valores del "4" al "9" no cumplen el mínimo de representación comentado anteriormente. Se ha optado por unificar todas las categorías mencionadas dentro de la categoría "3", por lo que de ahora en adelante dicha categoría representa el siguiente concepto: 3 o más personas dependientes. De esta forma se cuenta con la siguiente distribución de observaciones por categoría exhibida en la Tabla 7:

Tabla 7. Recategoricación de "Number of Dependants"

CATEGORÍAS DE "NUMBER OF DEPENDANTS"	OBSERVACIONES
0	5416
1	553
2	531
3+	543

Con la variable “Number of Referrals” ocurre algo similar, con la diferencia que una única categoría no cumple el mínimo necesario de representación, siendo esta la categoría “11”. Por lo tanto, esta categoría es incluida dentro de la categoría “10”, que de ahora en adelante representa “10 o más personas referidas a la compañía”.

6 Modelado de predicción de abandono de clientes

Para la construcción de modelos y su comparación posterior se han utilizado dos programas: SAS y Rstudio. El propósito del uso de más de un programa se debe a utilizar las diferentes configuraciones que permiten cada uno de estos y obtener diferentes variantes de los mismos algoritmos.

Para esta sección se realiza lo siguiente:

- Se han estandarizado las variables numéricas y se han transformado a “dummy” las variables categóricas
- Proceso de selección de variables: se han realizado pruebas con diferentes métodos para la selección de variables. En algunos algoritmos se ha utilizado este grupo de variables seleccionadas y en otros se ha permitido que sea el mismo algoritmo el que decida qué variable se debe utilizar, como en random forest o gradient boosting.
- Se crean los diferentes modelos utilizando todas las técnicas de machine learning mencionadas en la sección de metodología
    - Para el uso de Rstudio se transforma la variable objetivo (“Churn”) a valores “No” y “Yes”, ya que de esta forma se ejecuta correctamente el código.
- Se comparan a través de gráficos de boxplot en cuanto al área bajo la curva ROC y la tasa de fallos.

6.1 Selección de variables

Se han utilizado 4 procesos diferentes de selección de variables.

Los mismos son: selección con método “stepwise”, procedimiento “proc logistic” con selección “stepwise”, procedimiento “proc logistic” con “score” y el uso de la macro %randomselectlog provista por el profesor Portela (2020). Los resultados de cada uno de estos procesos se pueden encontrar en el Anexo I – Resultados de selección de variables.

Las variables que más apariciones han tenido en las diferentes repeticiones internas de cada uno de estos procesos y que coinciden en los cuatro mencionados anteriormente son:

- Monthly_Charge
- Online_Backup
- Online_Security
- Paperless_Billing
- Phone_Service
- Population
- Premium_Tech_Support
- Tenure_in_Months
- Total_Charges

6.2 Regresión Logística

Como primer modelo se utiliza la regresión logística, modelo que es utilizado como base de comparación para el resto de los modelos generados. A lo largo de la bibliografía comentada anteriormente, se utiliza la regresión logística como “benchmark” o base de comparación. Este algoritmo permite establecer si el aumento de complejidad de cálculo es valioso en comparación con los resultados obtenidos. Es decir, los resultados de la

regresión logística permiten verificar si el uso de modelos con mayor necesidad de procesamiento, como random forest o gradient boosting, es útil y genera una mejora en cuanto a la precisión en la predicción del abandono.

En Rstudio, con el uso de la función *cruzadalogistica* provista por Portela (2020), se generan dos regresiones logísticas: una con todas las variables y otra con algunas variables seleccionadas.

Figura 14.Tasa de fallos de modelos de regresión logística (R)

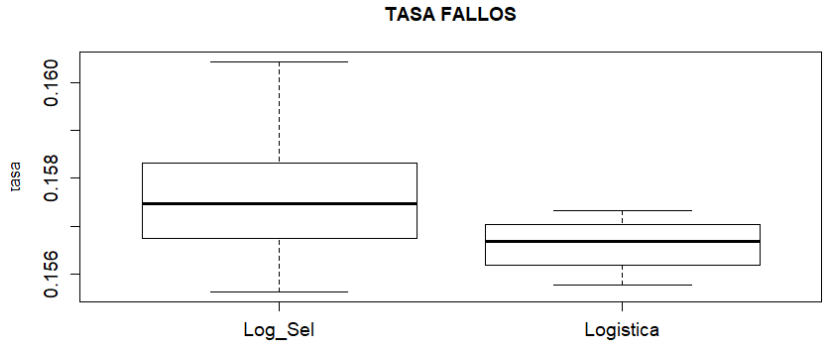
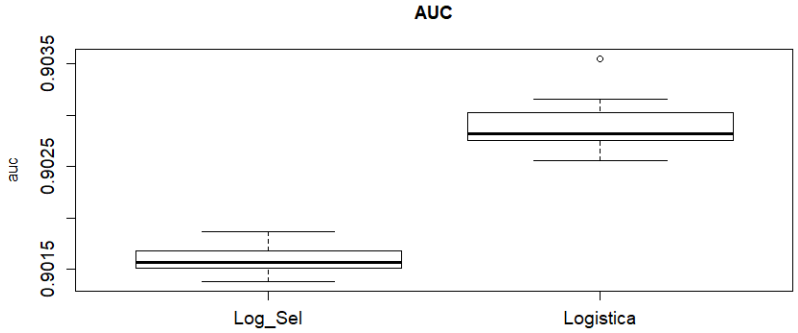


Figura 15. AUC Modelos de regresión logística (R)



Si se observan los valores que se han obtenido tanto para la tasa de fallos (Figura 14) como para el área bajo la curva ROC (Figura 15), los mismos han sido muy buenos. Se observa también que la pérdida de capacidad de predicción por el menor uso de variables es insignificante en comparación con el la ganancia en la velocidad del cálculo de cada uno de los modelos. Por otro lado, la varianza de la logística con selección aumenta en comparación a la que utiliza todas las variables.

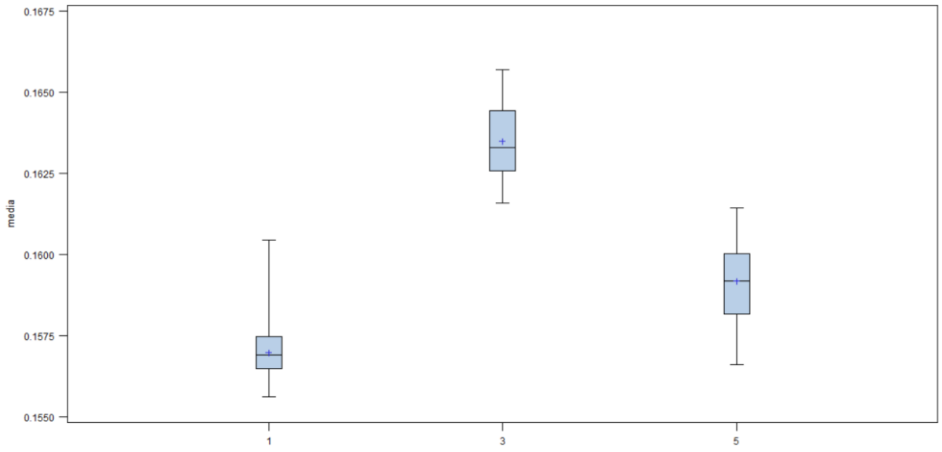
En SAS, se realiza una prueba con 3 regresiones logísticas, utilizando la macro *%cruzadalogistica*, las cuales se diferencian por la cantidad de variables que se le incluyen, determinadas en la Tabla 8:

Tabla 8. Regresiones logísticas (SAS)

Modelo	Logística 1	Logística 2	Logística 3
Variables	Todas	Monthly_Charge, Population, Tenure_in_Months, Total_Charges, Contract Married Number_of_Dependents Number_of_Referrals, Offer, Online_Backup, Online_Security, Paperless_Billing,	Population, Tenure_in_Months, Total_Charges, Monthly_Charge, Contract, Device_Protection_Plan, Internet_Type, Married, Number_of_Dependents, Number_of_Referrals Offer, Online_Backup, Online_Security, Paperless_Billing, Payment_Method, Phone_Service, Premium_Tech_Support,

		Payment_Method, Phone_Service	Streaming_Movies, Streaming_TV	Streaming_Music,

Figura 16. Tasa de fallos modelos de regresión logística (SAS)



El modelo 1, que utiliza todas las variables, es el que mejor se desempeña en cuanto a la tasa de fallos utilizando SAS, observando la Figura 16.

Se podrían utilizar los modelos de regresiones logísticas con todas las variables, tanto para R como para SAS, para comparar los demás modelos que se construyan, ya que ambos han obtenido una mediana de 15.70% de tasa de fallos, es decir, de casos mal clasificados, con una varianza baja en comparación a los otros modelos generados.

Tabla 9. Matriz de confusión y medidas de regresión logística con todas las variables (R)

	Referencia			
Prediction	**No**	**Yes**	**Accuracy**	0,8434
No	46845	6132	**Sensibilidad**	0,6719
Yes	4895	12558	**Especificidad**	0,9054

En este caso de regresión logística, observando los valores de la Tabla 9, con una tasa de corte de 0.5 para determinar si el cliente abandona o no, se establece que de los que abandonan se acierta un 67,19% de los casos, mientras de los que no abandonan se acierta un 90,54% de los casos.

6.3 Redes neuronales

Los principales hiperparámetros que se debe configurar en una red neuronal son el número de capas y el número de nodos que se utilizará. Debido a que el problema que se desarrolla en este trabajo no es de complejidad de redes profundas, o “Deep learning”, no es necesario el uso de más de una capa oculta. Por lo que en esta situación solo se debe establecer el número de nodos aproximado a utilizar.

Utilizando una fórmula provista por Portela (2020) se podría realizar este acercamiento a un número lógico de nodos a utilizar:

$$N \text{ parámetros} = h\,(k + 1) + h + 1$$

Siendo: h: nodos ocultos; k: variables input

Utilizando como valor mínimo de observaciones por nodo el número de 15, se obtienen los siguientes nodos exhibidos en la Tabla 10:

Tabla 10. Redes: número lógico de nodos

NODOS OCULTOS	3	5	7	9	11	13
NUMERO DE PARAMETROS	156	218	280	342	404	466
OBS. POR PARAMETRO	45,15	32,31	25,15	20,59	17,43	15,11

Para la creación de las diferentes redes neuronales en Rstudio, se utilizado el concepto de grilla con la librería *caret*, el cual permite combinar diferentes valores en los hiperparámetros de las redes para encontrar la que mejor se pueda adaptar a los datos.

Con el uso del *grid,* se han construido redes neuronales diferentes, con un número de 3, 5, 7, 9, 11 y 13 nodos, en combinación con *learning rates* de 0.1, 0.01 y 0.001.

En estos casos, el uso de un *decay* igual a 0,1 es el que mejor funciona para las diferentes redes creadas. En forma complementaria, se crean redes neuronales de similares características, con un conjunto de variables seleccionadas para ver qué resultados arrojan.

Utilizando la función *cruzadaavnnetbin,* se grafican los boxplot de cada una de las redes y se las compara con las regresiones logísticas:

Figura 17. Tasa de fallos de redes neuronales (R)

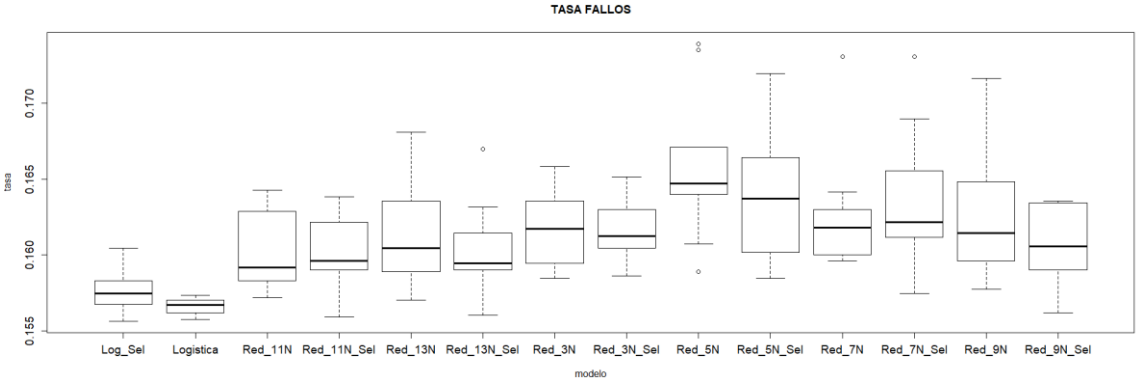
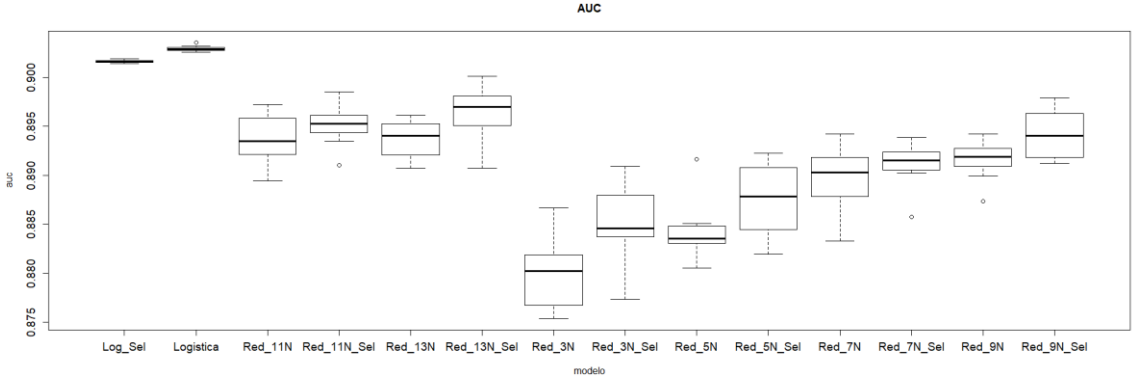


Figura 18. AUC de redes neuronales (R)



En el caso de las redes, la utilización de variables seleccionadas no parece ser la mejor de las alternativas, ya que, pese a aumentar el área bajo la curva ROC (Figura 18), aumenta la varianza de las mismas. Por lo que parece una mejor opción utilizar las redes con la totalidad de variables.

Aun así, la regresión logística obtiene un mejor rendimiento para estos datos, evidenciado tanto en la Figura 17 como en la Figura 18.

Tabla 11. Matriz de confusión y medidas de red neuronal 11 nodos (R)

	Referencia			
Prediction	**No**	**Yes**	**Accuracy**	0,8397
No	48636	8185	**Sensibilidad**	0,5621
Yes	3104	10505	**Especificidad**	0,9400

Con los resultados de la Tabla 11, en el caso de la red de 11 nodos creada en R, utilizando una tasa de corte de 0.5, se establece que de los que abandonan se acierta un 56,21% de los casos, mientras de los que no abandonan se acierta un 94% de los casos. Aunque el caso de la especificidad es muy buena la red, no clasifica correctamente los clientes de interés para este problema, es decir, aquellos que más probabilidades tienen de dejar la compañía, por lo que no resulta un modelo competitivo en comparación a la regresión logística.

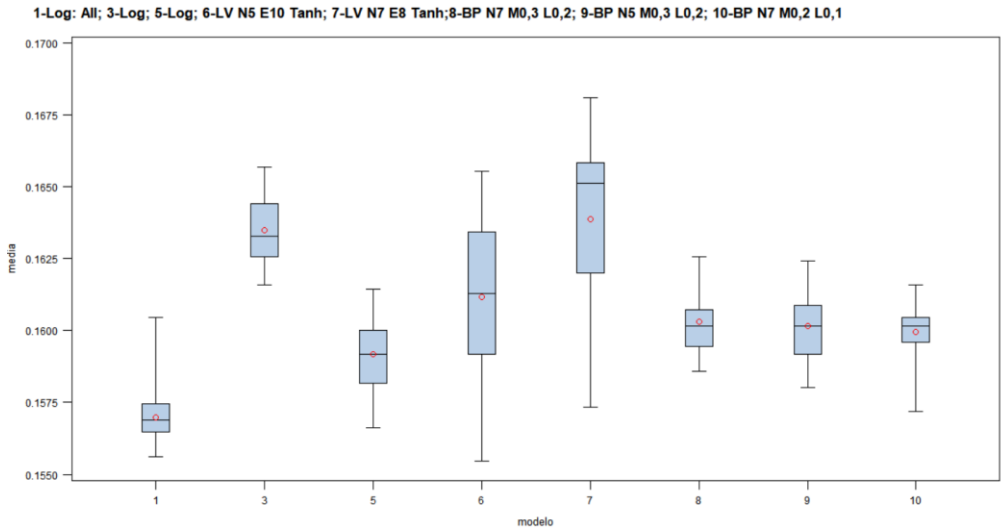
En SAS, se plantean diferentes modelos de redes, luego de un análisis de la configuración de los hiperpatámetros (Tabla 12), utilizando la macro *%cruzadabinarianeural*:

Tabla 12. Configuración de redes neuronales (SAS)

MODELOS	**6**	**7**	**8**	**9**	**10**
NODOS	5	7	7	5	7
ALGORITMO	Levmar	Levmar	Bprop	Bprop	Bprop
EARLY STOPPING	10	8	-	-	-
ACTIVACION	Tanh	Tanh	-	-	-
MOMENTUM	-	-	0.3	0.3	0.3
LEARNING RATE	-	-	0.2	0.2	0.2

Se grafican los boxplots de las tasas de fallo correspondientes a dichas redes y en comparación con las regresiones logísticas:

Figura 19. Tasa de fallos de redes neuronales (SAS)



Para este conjunto de datos es evidente, con los resultados de la Figura 19, que las redes neuronales pese a tener un buen rendimiento, tanto en R como en SAS, no son superiores al desempeño que logra la regresión logística. Esto se debe a que la mediana en la tasa de fallos de las redes no logra ser inferior a 15,90% y tienen una varianza mayor en cuanto a los resultados en comparación con la regresión logística, lo que genera que sean modelos no tan confiables.

6.4 Random forest y bagging

Se realiza una forma similar de aproximación que con las redes neuronales, utilizando la grilla de la librería *caret*.

Con la inclusión de la función *cruzadarfbin* se han generado 4 modelos de *random forest*. Complementando a los 4 modelos anteriores, se ha creado un modelo con utilizando todas las variables posibles, es decir, *bagging*. La parametrización de dichos modelos se ve reflejada en la Tabla 13.

Tabla 13. Configuración bagging y random forest (R)

	BAGGING	RF1	RF2	RF3	RF4
NODESIZE	10	10	10	10	20
MTRY	56	20	30	10	20
NTREE	100	600	600	600	600

No se especifica el número de datos a tomar en cada muestra, debido a que al utilizar validación cruzada se ha determinado que utilice todo el conjunto de datos de entrenamiento correspondiente a cada repetición.

Los resultados de tasa de fallos y área bajo la curva ROC se presentan a continuación:
Figura 20. Tasa de fallos bagging y random forest (R)

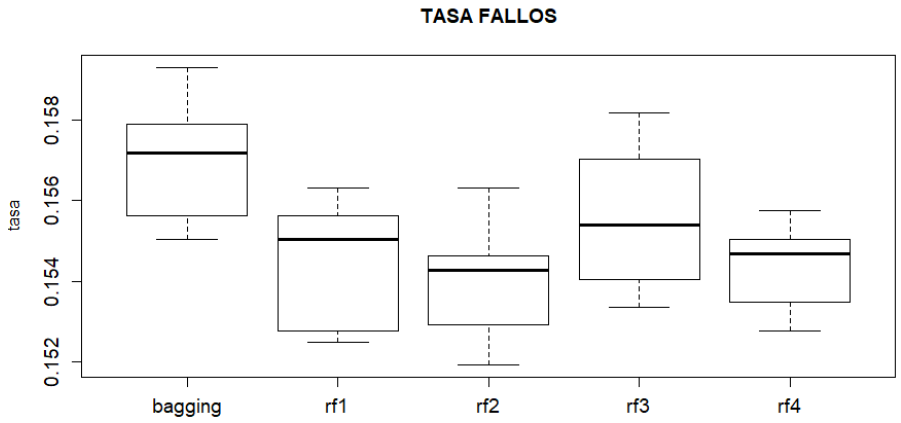
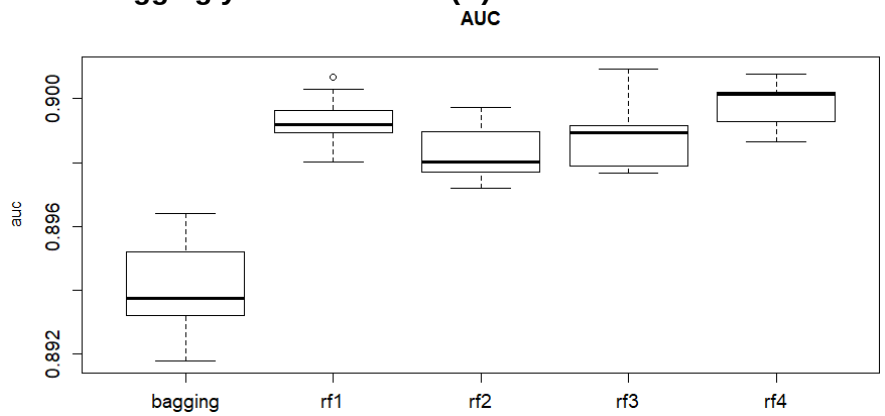


Figura 21. AUC bagging y random forest (R)



Como puede observarse en las Figura 20 y Figura 21, los 4 modelos de *random forest* se desempeñan bastante bien, pero por consistencia parece preferible el modelo "rf2" o el "rf4", debido a que no tienen valores por fuera de los límites del su gráfico de caja y sus varianzas son bajas. Entre ambos modelos, es mejor el "rf4", debido que su performance en cuanto a la tasa de fallos tiene una varianza menor que el "rf2".

Tabla 14. Matriz de confusión y medidas de rf4 (R)

	Referencia			
Prediction	**No**	**Yes**	**Accuracy**	0,8455
No	47703	6841	**Sensibilidad**	0,6340
Yes	4037	11849	**Especificidad**	0,9220

El modelo "rf4" ha tenido un desempeño similar al de la regresión logística, obteniendo una mejoría en la exactitud y en la especificidad, pero en deterioro de la sensibilidad, según los resultados de la Tabla 14.

6.5 Gradient boosting y Extreme Gradient Boosting

La bibliografía muestra que los algoritmos de gradient boosting suelen desempeñarse bastante bien en cuanto a la predicción de la probabilidad de abandono de clientes (Lemmens & Croux, 2006; Neslin et al., 2006; Tamaddoni et al., 2016).

Es por ello que se ha realizado un proceso extenso de búsqueda de los hiperparámetros óptimos para los algoritmos de gradient boosting y xgboost (extreme gradient boosting).

Ambos. procesos se encuentran descriptos en el Anexo II – Configuración de modelos de predicción de abandono.

Una vez que se han encontrado valores para los hiperparámetros que parecen adaptar estos dos algoritmos a los datos, se procede a utilizar las funciones de *cruzadagbmbin y cruzadaxgbmbin* para poder compararlos.

Configuración de Gradient Boosting:

- N.minobsnode = 30
- Shrinkage = 0,04
- N.trees = 500
- Interaction.depth = 2

La configuración de los diferentes modelos de XGboost se especifican en la Tabla 15.

Tabla 15. Configuración XGboost (R)

	XGBM	XGBM1	XGBM2	XGBM3	XGBM4	XGBM5
MIN_CHILD_WEIGHT	10	5	10	10	5	10
NROUNDS	800	1000	2500	2500	4000	800
SUBSAMPLE	1	1	1	0.8	0.8	1
ALPHA	3	6	3	3	6	0
LAMBDA	6	5	6	6	5	0
LAMBDA_BIAS	10	10	10	10	10	0
ETA (SHRINKAGE)	0.01	0.01	0.01	0.01	0.01	0.01
MAX_DEPTH	6	6	6	6	6	6

Los resultados se grafican a continuación:

Figura 22. Tasa de fallos de Gradient Boosting y XGboost (R)

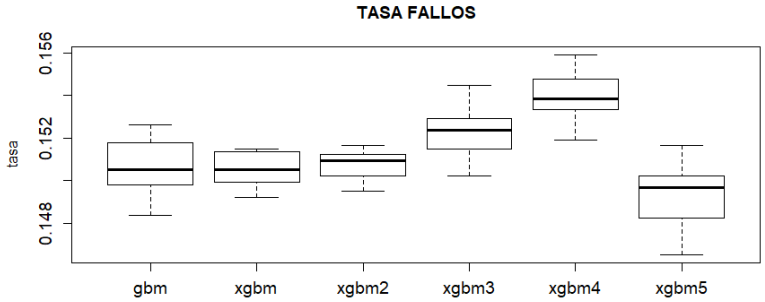
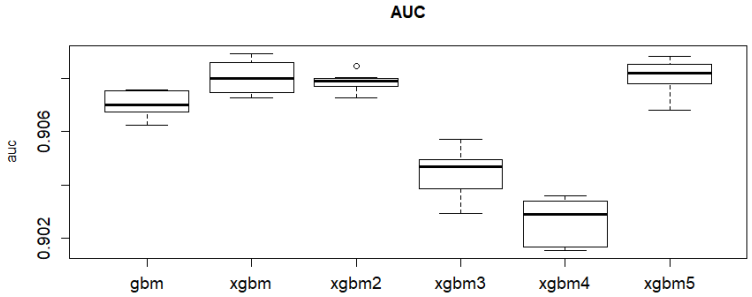


Figura 23. AUC de Gradient Boosting y XGboost (R)



Según los resultados exhibidos en las Figura 22 y Figura 23, el rendimiento de estos modelos generados es muy bueno, en cuanto al sesgo y la varianza.

Podría indicarse que, dentro de estos 6 modelos bajo análisis, el *xgbm* y el *xgbm2* son los que mejores resultados presentan, tanto en lo que concierne a la tasa de fallos y al área bajo la curva ROC. De todas formas, la diferencia (observando el margen de variación de tasa de fallos y de AUC) es muy baja entre todos los modelos en general.

Con el objetivo de destacar un ganador en este grupo de modelos, el *xgbm2* es el que menor variación de sesgo ha obtenido en todos los procesos de validación cruzada y el modelo que más consistente se desempeña con los datos en cuestión. Es por ello, que este sería el modelo elegido de este grupo.

Tabla 16. Matriz de confusión y medidas de XGboost2 (R)

	Referencia			
Prediction	**No**	**Yes**	**Accuracy**	0,8492
No	47466	6345	**Sensibilidad**	0,6605
Yes	4274	12345	**Especificidad**	0,9174

Según los datos del modelo xgbm2 de la Tabla 16, este modelo es levemente superior a la regresión logística: la sensibilidad es muy cercana y ha mejorado tanto la exactitud de las predicciones como el acierto en los casos en que el cliente no abandona.

6.6 Support Vector Machine

Se ha optado por realizar modelos con SVM en Rstudio, como otra alternativa para abordar el tema de predicción de probabilidad de abandono de clientes.

Luego de un proceso de calibración de los diferentes parámetros que permite modificar este algoritmo, se ha determinado la creación de los siguientes modelos con la función *cruzadaSVMbin:*

1. SVM Lineal: C igual a 0.07
2. SVM Polinomial: C igual a 0.01, grado igual a 2 y scale igual a 0.1
3. SVM RVF: C igual a 2 y sigma igual a 0.01

Los resultados se exponen a continuación:

Figura 24. Tasa de fallos de Support Vector Machine (R)

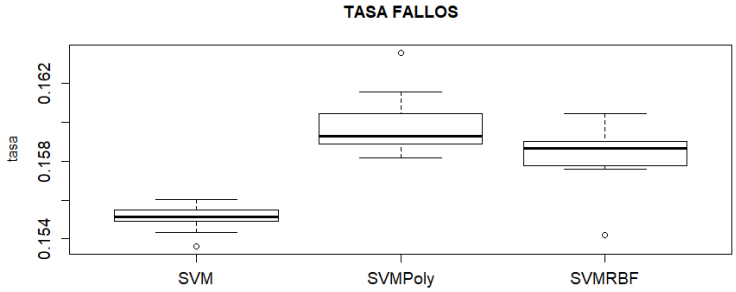
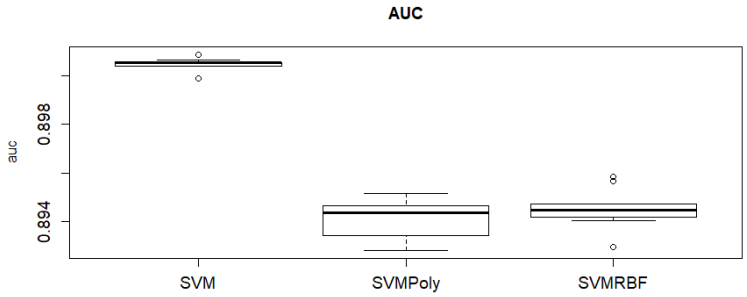


Figura 25. AUC de Support Vector Machine (R).



En base a las Figura 24 y Figura 25, el SVM lineal ofrece el mejor resultado comparándolo con los otros algoritmos de su misma familia. Pese a ser un algoritmo de mayor complejidad en cuanto a calculo que los demás anteriores, el SVM lineal ha obtenido buenos resultados en general.

Tabla 17. Matriz de confusión y medidas de SVM lineal (R)

	Referencia			
Prediction	**No**	**Yes**	**Accuracy**	0,8449
No	47233	6416	**Sensibilidad**	0,6567
Yes	4507	12274	**Especificidad**	0,9129

Con los valores de la Tabla 17, se puede decir que los resultados son competitivos en general para el SVM lineal, aunque no superan los obtenidos por el modelo *xgbm2* del apartado anterior, sobre todo en sensibilidad.

6.7 Ensamblado

Se ha optado por la creación de varios modelos ensamblando, utilizando los modelos anteriormente construidos.

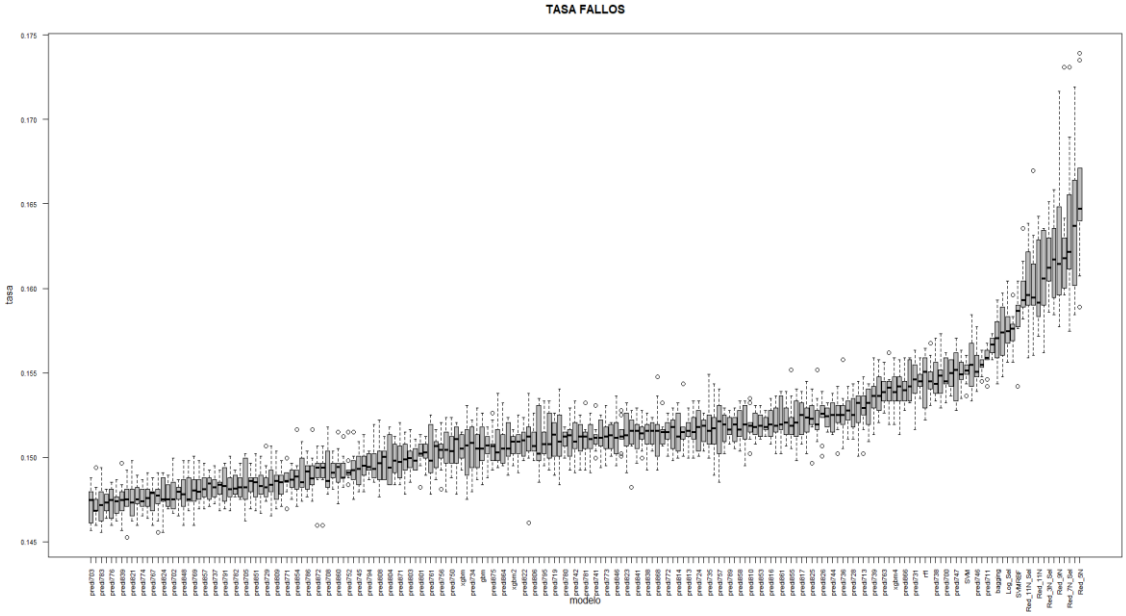
El propósito que se persigue es encontrar alguna combinación de modelos que permita una precisión más alta en la predicción del abandono de clientes y una disminución de la varianza.

Se han realizado varios modelos de ensamblado combinando todos los modelos generados en Rstudio, generando un total de 167:

- 55 modelos combinando 2 modelos individuales
- 106 modelos combinando 3 modelos individuales
- 3 modelos combinando 4 modelos individuales
- 3 modelos combinando 5 modelos individuales

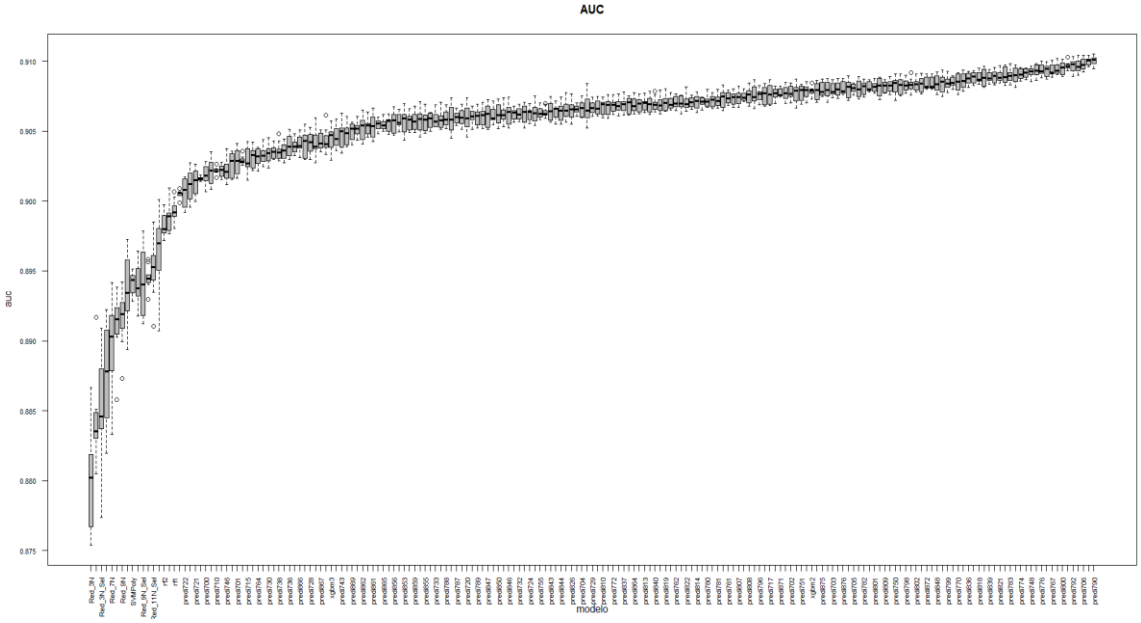
Una vez realizadas las combinaciones propuestas, se grafican los modelos ordenados por la mediana obtenida en la tasa de fallos:

Figura 26. Tasa de fallos de modelos ensamblados (R)



También se ordenan los modelos utilizando la mediana del AUROC:

Figura 27. AUC de modelos ensamblados (R)



Debido a que la cantidad de modelos generados, sumados a los individuales, representan una cantidad muy grande para que pueda ser visible en un gráfico (Figura 26 y Figura 27), se revisan los mejores 10 modelos en cuanto a tasa de fallos y AUC:

Tabla 18. Top10 modelos ensamblados según tasa de fallos (R)

	MODELO	TASA DE FALLO	MODELOS
1	predi703	0.147	logi + rf2
2	predi836	0.147	rf1 + xgbm2 + SVM
3	predi783	0.147	logi + rf3 + xgbm
4	predi775	0.147	logi + rf2 + xgbm
5	predi776	0.147	logi + rf2 + xgbm2
6	predi768	0.147	logi + rf1 + xgbm2

7	predi839	0.147	rf2 + xgbm2 + SVMRBF
8	predi842	0.148	rf3 + xgbm2 + SVM
9	predi821	0.148	rf2 + xgbm + SVM
10	predi784	0.148	logi + rf3 + xgbm2

Tabla 19. Top10 modelos ensamblados según AUC (R)

	MODELO	AUC	MODELOS
1	predi790	0.910	logi + gbm +xgbm
2	predi791	0.910	logi + gbm + xgbm2
3	predi706	0.910	logi + xgbm
4	predi797	0.910	logi + xgbm + xgbm4
5	predi792	0.910	logi + gbm + xgbm3
6	predi707	0.910	logi + xgbm2
7	predi800	0.909	logi + xgbm2 + xgbm4
8	predi775	0.909	logi + rf2 + xgbm
9	predi767	0.909	logi + rf1 + xgbm
10	predi793	0.909	logi + gbm + xgbm3

El único modelo que aparece en el "top 10", observando las Tabla 18 y Tabla 19, es el denominado "predi775", originado por la combinación de los modelos *regresión logística*, *random forest 2* y *extreme gradient boosting*.

Aun así, los modelos "predi706" y "predi707", construidos con *regresión logística* y *extreme gradient boosting* son competitivos.

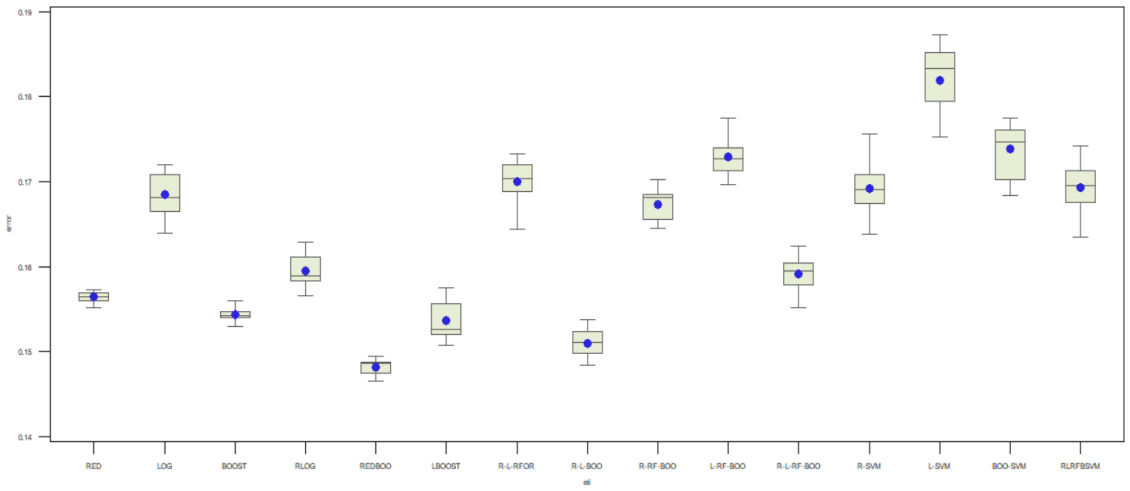
En general, la combinación de *regresión logística* con los algoritmos de *gradient boosting* y *extreme gradient boosting* parecen ser complementarios en cuanto a la predicción de la variable objetivo, debido a que son las combinaciones de algoritmos que más se repiten en las tablas Tabla **18** y Tabla **19**.

En SAS, se utiliza la macro *%cruzadastackcon*, para crear los siguientes modelos en forma individual, que corresponden a los mejores modelos de sus respectivos grupos, y su combinación para construir diferentes modelos ensamblados:

- Regresión logística con todas las variables
- Red neuronal utilizando el algoritmo de Backpropagation de 7 nodos, 0.3 de momentum y learning rate de 0.2
- Random Forest: 150 árboles, 8 variables a seleccionar, tamaño de hoja de 25 observaciones y una profundidad de 8 niveles máxima
- Gradient Boosting: 200 iteraciones, shrinkage de 0.01, leafsize de 25, maxbranch de 4 y maxdepth de 8
- SVM lineal con el paramétro C igual a 10

El resultado obtenido de la macro anterior es el siguiente:

Figura 28. Tasa de fallos de modelos individuales y ensamblados (SAS)



Se han excluido los modelos de random forest, SVM lineal y de su combinación debido a que no permitía observar con claridad las diferencias del resto de los modelos.

Dentro de la Figura 28, como modelos individuales se destacan tanto la *red neuronal* y el *gradient boosting*. Como modelos ensamblados son destacables: la combinación de los dos individuales mencionados anteriormente (*REDBOO*), *LBOOST* y *R-L-BOO*.

El caso del modelo *REDBOO* es el único cuyos errores de predicción ha estado por debajo del 15% en la totalidad de las 10 repeticiones de validación cruzada que se han realizado con variación de semilla en SAS.

6.8 Elección del Mejor modelo

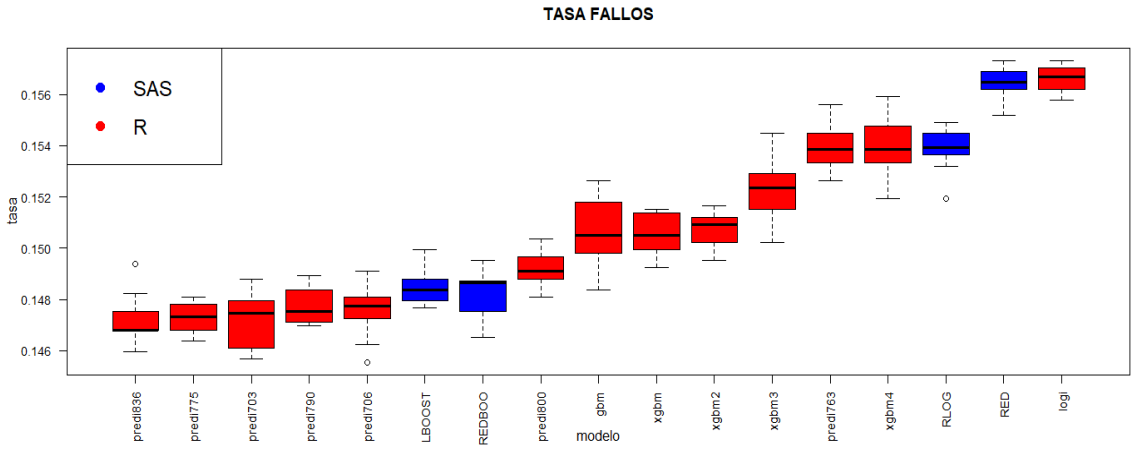
En general, el rendimiento de las redes neuronales ha sido mejor en SAS que en R. Ocurre una situación diferente para el resto de los modelos, los cuales han obtenido una mejor performance utilizando los paquetes y variantes que ofrece R.

Cabe destacar el buen rendimiento de *gradient boosting* y *extreme gradient boosting* para los datos que se disponen. También la *regresión logística* podría utilizarse, ya que ha obtenido muy buenos resultados, competitivos versus algoritmos más complejos.

En el caso de los modelos ensamblados, se obtienen mejores resultados en algunos casos que se combinan algoritmos que detectan mejor que otros ciertas secciones de los datos, y en la unión con otros generan una “sinergia” positiva en cuanto a la predicción del abandono de cliente.

A continuación, se realiza una comparación gráfica de la tasa de fallos de los mejores modelos obtenidos en R y SAS:

Figura 29. Tasa de fallos de mejores modelos (R y SAS)



Como comentarios finales referidos a los modelos, en el caso que se desee un modelo con mayor interpretabilidad, dejando en segundo lugar rendimiento, la *regresión logística* puede ser utilizada para este caso.

En otro caso, en el cual la precisión de las predicciones sea lo más importante, dejando de lado la interpretación del modelo, lo mejor sería utilizar *gradient boosting* o *xgbm2*, o algún modelo ensamblado, como el *Red-Boost* o *LBoost* en SAS o el modelo "predi775" en R (*regresión logística*, *random forest 2* y *extreme gradient boosting*), según las comparaciones observadas en la Figura 29.

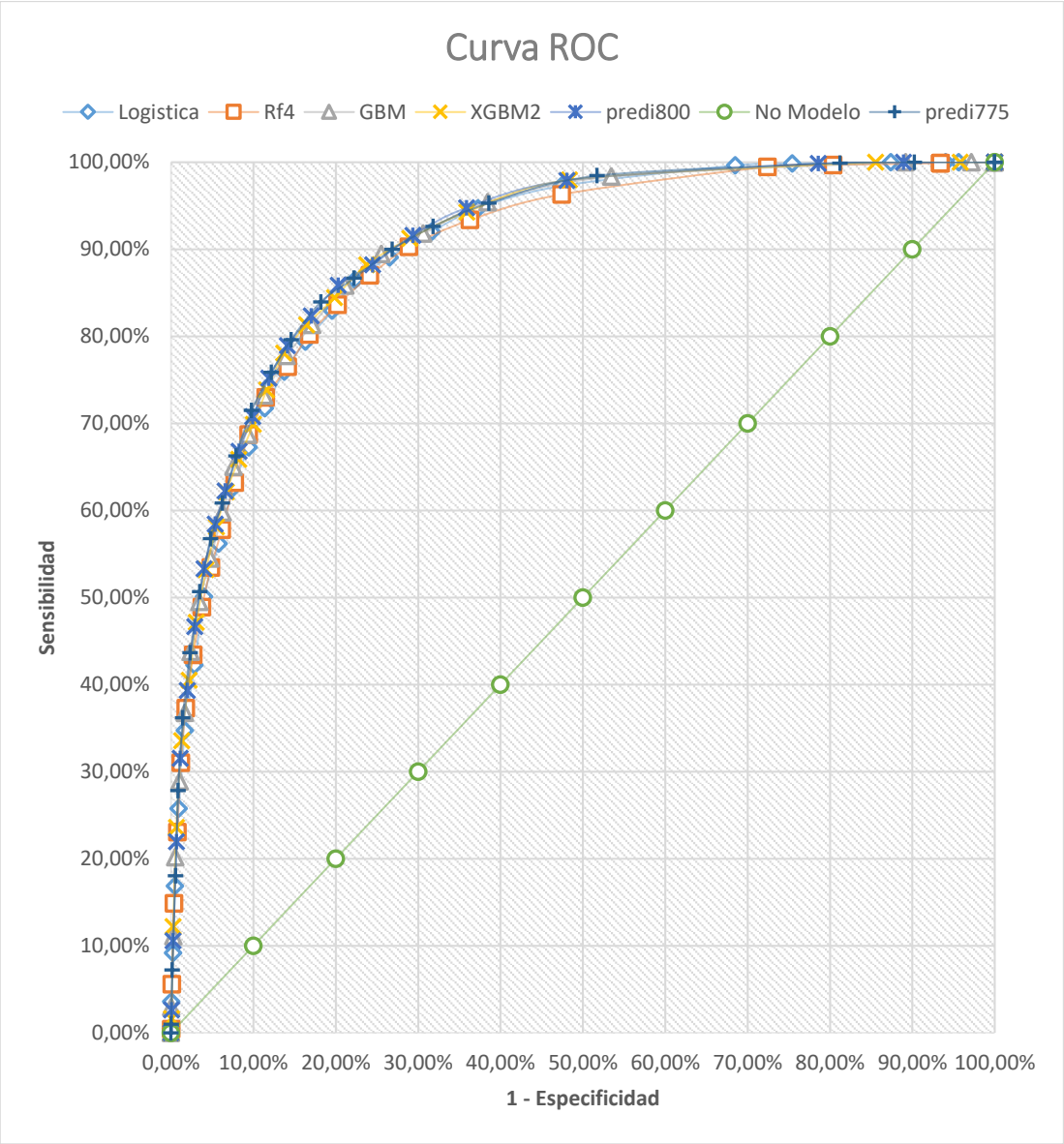
6.8.1 Análisis de tasas de corte

Se ha realizado un gráfico (Figura 30), utilizando diferentes tasas de corte, para establecer qué modelo es mejor de los anteriormente mencionados.

A continuación, se presentan las curvas ROC de los siguientes modelos:

- Regresión logística
- Random forest (rf4)
- Gradient boosting (gbm)
- Extreme gradient boosting (xgbm2)
- Modelo "predi775": ensamblado de regresión logística, random forest (rf2) y extreme gradient boosting (xgbm)
- Modelo "predi800": ensamblado de regresión logística y dos extreme gradient boosting (xgbm2 y xgbm4)
- "No modelo": modelo "aleatorio".

Figura 30. Curva ROC de mejores modelos

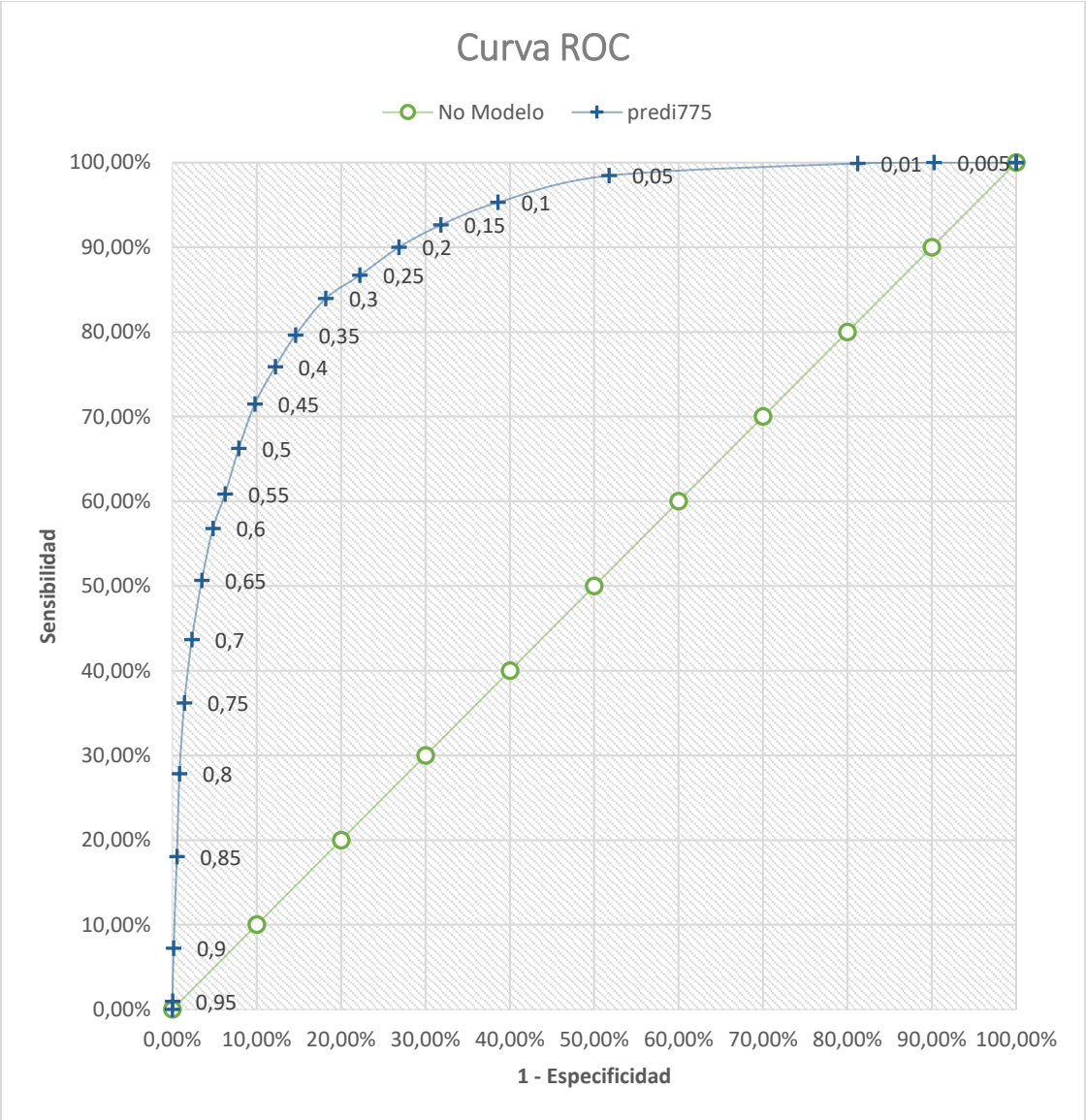


En general los modelos han obtenido resultados parecidos, pero se pueden destacar algunas cuestiones:

- Random forest (rf4) no genera los mejores resultados en comparación al resto
- La regresión logística obtiene resultados muy buenos
- Se aprecia que el modelo “predi775” es el que mejor se desempeña, aunque no muy alejado del resto de los modelos contruidos

Si se revisa del modelo “predi775” las tasas de cortes para cada punto de la curva ROC representada anteriormente se puede apreciar lo siguiente:

Figura 31. Tasa de corte de modelo "predi775"



Según la Figura 31, para este conjunto de datos el valor máximo de área bajo la curva ROC parece encontrarse en la tasa de corte de 0,3 del modelo “predi775”. En dicho punto el valor de sensibilidad es de 83,95% y el valor de especificidad es de 81,81%.

Esto significa que utilizando este valor de tasa de corte el modelo “predi775” ha podido clasificar correctamente a los clientes que abandonan en el 83,95% de los casos y que ha detectado correctamente los clientes que no abandonan en un 81,81%.

6.9 Importancia de variables

Además de haber realizado un proceso de pruebas y construcción de modelos con diferentes algoritmos también es importante verificar qué variables son utilizadas por estos y si existe algún grupo de estas que se repitan o que tomen un papel de mayor relevancia para la predicción de abandono de clientes.

Para realizar este proceso se toman las variables que más relevancia tienen de 4 modelos generados en Rstudio: regresión logística (Tabla 20), gradient boosting (Tabla 22), extreme gradient boosting o “xgbm2” (Tabla 23) y random forest o “rf4” (Tabla 21).

Tabla 20. Importancia de variables regresión logística

VARIABLE	REL.INF	STD. ERROR	PR(>|Z|)
TENURE_IN_MONTHS	13,5069	0,04795	0,00000
NUMBER_OF_REFERRALS_1	9,6200	0,03680	0,00004
NUMBER_OF_DEPENDENTS_0	7,3438	0,17079	0,00000
PAYMENT_METHOD_CREDIT_CARD	2,6099	0,60491	0,00000
PREMIUM_TECH_SUPPORT	0,4556	0,15167	0,00000
STREAMING_MOVIES	0,4093	0,18663	0,00000
CONTRACT_ONE_YEAR	0,3869	0,19133	0,00000
STREAMING_TV	0,1925	0,57932	0,00023
NUMBER_OF_REFERRALS_8	0,1397	0,55822	0,00003
AVG_MONTHLY_LONG_DISTANCE_CHARGE	0,1391	1,02758	0,00002
CLTV	0,1250	0,15438	0,00003
INTERNET_TYPE_CABLE	0,1172	0,15694	0,00000
NUMBER_OF_REFERRALS_9	0,0852	0,13548	0,00020
ONLINE_BACKUP	0,0233	0,22100	0,00000
NUMBER_OF_DEPENDENTS_1	0,0000	0,13393	0,00000
NUMBER_OF_DEPENDENTS_2	0,0000	0,13645	0,00000
NUMBER_OF_REFERRALS_2	0,0000	0,08337	0,00004
NUMBER_OF_REFERRALS_3	0,0000	0,43974	0,00000
NUMBER_OF_REFERRALS_4	0,0000	0,13787	0,00000
NUMBER_OF_REFERRALS_5	0,0000	0,25284	0,00000
UNLIMITED_DATA	0,0000	0,21938	0,00032
PAPERLESS_BILLING	0,7900	0,10833	0,00711
STREAMING_MUSIC	0,4583	0,19070	0,00780
PHONE_SERVICE	0,0000	0,14750	0,00917
AVG_MONTHLY_GB_DOWNLOAD	2,1944	0,18451	0,01152
TOTAL_CHARGES	0,7815	0,28305	0,04412
TOTAL_LONG_DISTANCE_CHARGES	0,5686	0,19786	0,03690
MULTIPLE_LINES	0,0440	0,24073	0,04566
NUMBER_OF_REFERRALS_7	0,0368	0,13600	0,03196

Tabla 21. Importancia de variables de modelo rf4

VARIABLE	MEAN DECREASE ACCURACY	MEAN DECREASE GINI
CONTRACT_MONTH_TO_MONTH	126,77	395,09
NUMBER_OF_REFERRALS_1	67,68	100,31
AGE	60,68	151,21
NUMBER_OF_DEPENDENTS_0	52,83	81,31
MONTHLY_CHARGE	43,50	157,32
TENURE_IN_MONTHS	36,74	185,85
TOTAL_CHARGES	35,01	116,47
TOTAL_LONG_DISTANCE_CHARGES	29,54	93,92
PAYMENT_METHOD_CREDIT_CARD	24,53	33,43
AVG_MONTHLY_GB_DOWNLOAD	24,44	71,52
INTERNET_TYPE_FIBER_OPTIC	21,99	89,75
PREMIUM_TECH_SUPPORT	17,94	11,62
NUMBER_OF_REFERRALS_0	17,23	18,59

NUMBER_OF_REFERRALS_9	17,20	4,45
AVG_MONTHLY_LONG_DISTANCE_CHARGE	16,54	64,87
MARRIED	15,59	12,67
ONLINE_SECURITY	15,28	13,70
STREAMING_MUSIC	13,27	12,89
CLTV	12,42	70,35
NUMBER_OF_REFERRALS_8	12,35	3,01
NUMBER_OF_DEPENDENTS_1	12,01	4,48
INTERNET_TYPE_DSL	11,16	4,99
OFFER_OFFER_B	10,93	6,00
STREAMING_TV	10,23	7,50
UNLIMITED_DATA	10,19	7,99
PAPERLESS_BILLING	10,16	12,42
PAYMENT_METHOD_BANK_WITHDRAWAL	9,03	7,55
STREAMING_MOVIES	8,99	5,62
INTERNET_TYPE_CABLE	8,13	4,47
MULTIPLE_LINES	8,10	6,24
ONLINE_BACKUP	6,54	5,84
NUMBER_OF_REFERRALS_7	6,51	1,59
NUMBER_OF_DEPENDENTS_2	6,34	2,78
NUMBER_OF_REFERRALS_6	6,30	1,62
DEVICE_PROTECTION_PLAN	6,29	3,92
POPULATION	5,88	77,67
TOTAL_EXTRA_DATA_CHARGES	5,85	18,82
NUMBER_OF_REFERRALS_4	5,29	1,80
PHONE_SERVICE	3,49	1,92
OFFER_NONE	2,89	5,74
CONTRACT_ONE_YEAR	2,59	14,34
NUMBER_OF_REFERRALS_2	2,03	2,85
TOTAL_REFUNDS	1,38	8,33
OFFER_OFFER_D	1,18	2,07
NUMBER_OF_REFERRALS_5	0,68	2,06
OFFER_OFFER_A	0,42	0,85
NUMBER_OF_REFERRALS_3	-0,17	2,51
GENDER_FEMALE	-2,24	4,63
OFFER_OFFER_C	-2,40	2,17

Tabla 22. Importancia de variables modelo gbm

VARIABLE	REL.INF
CONTRACT_MONTH_TO_MONTH	38,0700
TENURE_IN_MONTHS	13,5069
NUMBER_OF_REFERRALS_1	9,6200
AGE	8,0959
NUMBER_OF_DEPENDENTS_0	7,3438
MONTHLY_CHARGE	5,3931
INTERNET_TYPE_FIBER_OPTIC	4,1764
PAYMENT_METHOD_CREDIT_CARD	2,6099
AVG_MONTHLY_GB_DOWNLOAD	2,1944
NUMBER_OF_REFERRALS_0	2,0790
POPULATION	1,1924
ONLINE_SECURITY	0,8910

PAPERLESS_BILLING	0,7900
TOTAL_CHARGES	0,7815
TOTAL_LONG_DISTANCE_CHARGES	0,5686
STREAMING_MUSIC	0,4583
PREMIUM_TECH_SUPPORT	0,4556
STREAMING_MOVIES	0,4093
CONTRACT_ONE_YEAR	0,3869
STREAMING_TV	0,1925
NUMBER_OF_REFERRALS_8	0,1397
AVG_MONTHLY_LONG_DISTANCE_CHARGE	0,1391
CLTV	0,1250
INTERNET_TYPE_CABLE	0,1172
NUMBER_OF_REFERRALS_9	0,0852
OFFER_OFFER_B	0,0493
MULTIPLE_LINES	0,0440
NUMBER_OF_REFERRALS_7	0,0368
ONLINE_BACKUP	0,0233
TOTAL_EXTRA_DATA_CHARGES	0,0083
TOTAL_REFUNDS	0,0076
OFFER_NONE	0,0057
NUMBER_OF_REFERRALS_6	0,0031

Tabla 23. Importancia de variables modelo xgbm2

VARIABLE	OVERALL
CONTRACT_MONTH_TO_MONTH	0.341073
AGE	0.089374
TENURE_IN_MONTHS	0.069814
NUMBER_OF_REFERRALS_1	0.061416
NUMBER_OF_DEPENDENTS_0	0.059444
MONTHLY_CHARGE	0.056355
POPULATION	0.033383
CLTV	0.031745
TOTAL_LONG_DISTANCE_CHARGES	0.027862
TOTAL_CHARGES	0.027312
AVG_MONTHLY_LONG_DISTANCE_CHARGE	0.023197
PAYMENT_METHOD_CREDIT_CARD	0.022701
NUMBER_OF_REFERRALS_0	0.022600
AVG_MONTHLY_GB_DOWNLOAD	0.020642
INTERNET_TYPE_FIBER_OPTIC	0.015658
STREAMING_MUSIC	0.010884
CONTRACT_ONE_YEAR	0.010857
ONLINE_SECURITY	0.010031
PAPERLESS_BILLING	0.008054
PREMIUM_TECH_SUPPORT	0.007253

Con el propósito de conocer qué variables aparecen más veces en estos modelos y en una mejor posición se realiza el siguiente proceso:

- Se asignan posiciones a cada una de las variables por cada uno de los modelos
- Se cuenta en cuantos modelos aparece cada una de las variables
- Solamente se contemplan variables que aparezcan como mínimo en tres de los cuatro modelos presentados

- Se realiza un cálculo de posición ponderada en el cual se suman las posiciones de la variable en cada uno de los modelos y se divide por la cantidad de modelos en los que aparece

Como resultado del proceso anterior se obtiene una tabla de importancia general de variables, exhibido en la Tabla 24.

Tabla 24. Importancia de variables ponderada

VARIABLE	LOGI	RF4	GBM	XGBM2	POSICIÓN PONDERADA
CONTRACT_MONTH_TO_MONTH		1	1	1	1,00
NUMBER_OF_REFERRALS_1	2	2	3	4	2,75
TENURE_IN_MONTHS	1	6	2	3	3,00
AGE		3	4	2	3,00
NUMBER_OF_DEPENDENTS_0	3	4	5	5	4,25
MONTHLY_CHARGE		5	6	6	5,67
PAYMENT_METHOD_CREDIT_CARD	4	9	8	12	8,25
INTERNET_TYPE_FIBER_OPTIC		11	7	15	11,00
NUMBER_OF_REFERRALS_0		13	10	13	12,00
PREMIUM_TECH_SUPPORT	5	12	17	20	13,50
TOTAL_CHARGES	26	7	14	10	14,25
AVG_MONTHLY_GB_DOWNLOAD	25	10	9	14	14,50
AVG_MONTHLY_LONG_DISTANCE_CHARGE	10	15	22	11	14,50
TOTAL_LONG_DISTANCE_CHARGES	27	8	15	9	14,75
CLTV	11	19	23	8	15,25
ONLINE_SECURITY		17	12	18	15,67
NUMBER_OF_REFERRALS_8	9	20	21		16,67
NUMBER_OF_REFERRALS_9	13	14	25		17,33
STREAMING_MOVIES	6	28	18		17,33
STREAMING_TV	8	24	20		17,33
POPULATION		36	11	7	18,00
STREAMING_MUSIC	23	18	16	16	18,25
PAPERLESS_BILLING	22	26	13	19	20,00
CONTRACT_ONE_YEAR	7	41	19	17	21,00
INTERNET_TYPE_CABLE	12	29	24		21,67
ONLINE_BACKUP	14	31	29		24,67
NUMBER_OF_DEPENDENTS_1	15	21	42		26,00
MULTIPLE_LINES	28	30	27		28,33
NUMBER_OF_REFERRALS_7	29	32	28		29,67
NUMBER_OF_DEPENDENTS_2	16	33	43		30,67
UNLIMITED_DATA	21	25	49		31,67
NUMBER_OF_REFERRALS_2	17	42	44		34,33
NUMBER_OF_REFERRALS_4	19	38	46		34,33
NUMBER_OF_REFERRALS_3	18	47	45		36,67
PHONE_SERVICE	24	39	48		37,00
NUMBER_OF_REFERRALS_5	20	45	47		37,33

Observando la tabla de resultados de relevancia de las variables (Tabla 24) se puede detectar que las primeras 7 variables ocupan los primeros diez lugares de los modelos en los que aparecen, por lo que podría interpretarse que son las más útiles en cuanto a la predicción de abandono de clientes.

Analizando las primeras 7 variables en comparación con la variable objetivo (“Churn”) se pueden establecer las siguientes afirmaciones:

- La mayoría de los clientes que abandonan tienen un contrato de tipo mes a mes (Month _to_Month).
- A partir de que el cliente ha referido al menos una persona la probabilidad de abandono es menor (Number_of_referrals_1).
- A mayor cantidad de meses que el cliente permanece en la empresa su probabilidad de abandono es menor (Tenure_in_Months).
- A medida que el cliente aumenta en edad su probabilidad de abandono aumenta (Age).
- En caso que el cliente no tenga personas a cargo su probabilidad de abandono aumenta (Number_of_dependents_0).
- En general, cuanto más altos son los cargos mensuales es más probable que el cliente abandone (Monthly_charge).
- Los clientes que realizan el pago de los servicios contratados a través de tarjeta de crédito tienen una tendencia mucho menor, casi la mitad, que el resto de los clientes que realizan sus pagos mediante otros medios (Payment_method_credit_card).

Se aclara que las afirmaciones anteriores son realizadas en base a los datos con los que se ha realizado el presente trabajo. Con una mayor cantidad de datos de la misma empresa, por ejemplo, de otro trimestre, podría establecerse una mayor precisión en cuanto a la declaración de la importancia de las variables y su origen.

7 Segmentación de clientes

7.1 Introducción

¿Por qué es importante segmentar? ¿Es sencillo personalizar individualmente un servicio? ¿Es rentable? ¿Es beneficioso tratar a todos los clientes por igual? ¿Existen clientes que se parecen?

La predicción de abandono de cliente, o la probabilidad de abandono de cada uno de los clientes de la empresa, es de vital importancia para la subsistencia de la misma. Aun así, la utilización de los modelos obtenidos de forma única no es la mejor opción, ni la más eficiente ni eficaz.

Es posible que el cliente con mayor probabilidad de abandono no sea el más rentable de retener ni el que menos recursos implicará para dicha acción. También podría ocurrir que los clientes más valiosos, en cuanto a rentabilidad y posibilidad de generación de beneficios, no sean los de mayor probabilidad de abandono, pero si uno de estos lo fuera sería óptimo intentar de retenerle cuanto antes.

Es por ello que los modelos obtenidos deben ser complementados por una correcta segmentación de clientes, para poder enfocar los recursos escasos de la organización en retener aquellos clientes más importantes y valiosos para esta.

Los clientes pueden ser clasificados en diferentes grupos, dependiendo del comportamiento de compra, rango etario, frecuencia de uso o visita a tiendas, gasto promedio, actitudes, estilo de vida y otras características de los mismos.

¿Por qué es tan importante segmentar la base de clientes? Algunos beneficios son:

- Distinción de tipos de clientes: ofrecer el mismo servicio y producto a toda la base de clientes no es óptimo.
- Foco en los que más importan: agrupar clientes y entender cuáles tienen mayor valor potencial para la empresa permite clarificar la dirección estratégica de la misma.
- Descubrimiento de similitudes y posibilidad de ofrecimientos a grupos con características similares: no es necesario en todos los sectores adaptar los servicios hasta el mínimo detalle, es decir, personalizarlo completamente. Una agrupación inteligente de estos permite diferentes variantes a ofrecer de utilidad.
- Adaptación de productos y servicios: con pequeñas modificaciones y ofrecimientos es posible aumentar la satisfacción de los diferentes grupos de clientes, aun con el mismo servicio o producto base.
- Mejora de productos y servicios: con la introducción de mejoras al servicio o a los productos, se puede obtener una respuesta de los clientes en cuanto al valor percibido a las mismas. Recabar las opiniones de los distintos clientes en cuanto a lo ofrecido permite mejorar los productos y servicios en cuestiones que aportan valor para los consumidores, que permite cobrar un precio diferencial por ello.
- Optimización de estrategia de precios: un factor crucial en la mayoría de los sectores es el precio que se debe pagar por un servicio o producto. Saber diferenciar qué clientes pagarán más por una mejora en el mismo puede mejorar la relación con estos, centrándola en lo que el cliente necesita.

- Modos de comunicación: algunas personas pueden preferir una comunicación más informal que otras; por otro lado, los medios de comunicación pueden ser muy variados: persona a persona, contacto telefónico, correos electrónicos, mensajes al móvil y demás. Saber de qué modo acercarse al cliente y bajo qué características se puede hacerlo es vital para una empresa, ya que mejorará sustancialmente la relación con estos y permitirá aumentar la precisión del uso de recursos.
- Lanzamiento de nuevos productos y servicios: conocer qué cliente, cuándo realizó una compra o de qué forma utiliza los servicios genera un gran conjunto de datos de alto valor para una empresa. Ante la aparición de nuevos servicios o productos dentro de la misma, se puede dirigir una campaña de venta a los clientes que mayor probabilidad de contratación o compra tengan sobre estos, basándose en la información histórica de los mismos. Incluso para clientes relativamente nuevos es posible realizar un proceso similar conociendo el grupo o segmento al que pertenece, ya que podría utilizar información de otros clientes de mayor antigüedad con rasgos similares.
- Organización interna de la compañía: si existen segmentos suficientemente rentables e interesantes para la organización, la misma puede adaptar su estructura, procesos, departamentos y puestos en base a estos grupos de clientes para enfocar sus recursos con una mayor precisión.

“Si una campaña de marketing tiene como objetivo retener a los clientes existentes, el sistema CRM de la compañía es central para poder analizar los registros de comportamiento de compra y el perfil de cada uno de estos, lo que permitirá, al personal de marketing, ajustar la oferta hacia las necesidades insatisfechas de los mismos. Se pueden realizar diferentes análisis, como por ejemplo: segmentación por ciclo de vida del cliente, generación de perfiles y análisis RFM” (V. Kumar, 2018)

7.2 Segmentación

Como complemento a los modelos obtenidos de predicción de ”churn” se utiliza el algoritmo de k-means, o k-medias, sobre el conjunto de datos para detectar grupos de clientes que guarden características similares. La idea principal es agrupar clientes que sean homogéneos entre sí, con el objetivo de poder aplicar acciones de retención sobre estos que puedan ser útiles para más de un individuo.

En principio, la decisión de cuál es el valor de *k* puede ser determinado por el área de marketing en una compañía, el cual puede estar condicionado por el número de segmentos que la empresa desea gestionar o que puede abordar con los recursos que dispone. Aunque también es viable aplicar algún método de k óptimo para descubrir cual es el mejor número de grupos a utilizar en este conjunto de datos.

Para la aplicación de este algoritmo es importante contar con las variables numéricas estandarizadas. Esto se debe a que las variables con mayores valores puedan prevalecer sobre el resto y esto no implica necesariamente que sean las más útiles para dividir a los clientes en segmentos.

Se han realizado diferentes pruebas en cuanto a la búsqueda de un valor de k óptimo para los clientes de esta empresa (ver Anexo III – Valor óptimo de k). En general, la mayoría de los resultados oscilan en valores entre 3 y 4 clusters.

A fines prácticos se utilizan 4 clusters para poder dividir a los clientes en diferentes segmentos.

En un ámbito empresarial lo importante de esta clusterización es que, una vez validada en cuanto a la interpretación y validez de los grupos de clientes, la misma sea incluida en el sistema CRM de la compañía para poder realizar diferentes acciones comerciales como: campañas de desarrollo, de retención, programas de fidelización y ofrecimiento de nuevos servicios y/o descuentos.

Con la ayuda de una herramienta de inteligencia de negocios, Power BI, se ha trasladado la información de los datos de clientes y la asignación de cada uno de estos al clúster correspondiente para poder visualizar sus características.

Figura 32. Clusters de Clientes

Cluster	Clientes	Mín. Edad	Promedio Edad	Máx. Edad	Prom. Dependants	Prom. Referrals	Prom. Charges	Prom. GB Download	Prom. Long Dist Charges
1	3005	19	47,51	80	0,33	1,15	71,18	26,80	21,01
2	319	19	48,62	80	0,50	2,19	68,38	20,29	21,18
3	1964	19	47,25	80	0,50	3,02	89,90	27,31	27,17
4	1755	19	43,58	80	0,67	2,08	24,98	2,19	21,90

Como puede observarse en la Figura 32, no existen diferencias de los clientes en cuanto a la edad entre los diferentes grupos, si se observa el mínimo, el promedio y el máximo de edad por cluster.

Las diferencias más destacables de la Figura 32 entre los diferentes “clusters” son: el CLTV (calculado por el CRM), antigüedad como clientes (Tenure), personas referidas a la empresa (Prom. Referrals), el uso de datos e internet (Prom. GB Download) y la contratación de servicios de streaming (movies, music y TV).

8 Acciones comerciales

No todos los clientes en una empresa son igualmente valiosos ni atractivos comercialmente.

No sería optimo intentar retener aquellos clientes que presenten una mayor probabilidad de abandono, debido a que puede que no sean los clientes más rentables para la compañía o simplemente no se encuentran entre sus objetivos estratégicos mantener ciertos perfiles de consumidores.

Entonces, es necesario aplicar alguna forma de priorización para realizar acciones comerciales de retención. Es por ello, que combinando las probabilidades de abandono obtenidas con el mejor modelo de predicción y ordenando a los clientes por los segmentos a los que pertenecen, es posible establecer esta priorización.

La combinación de este algoritmo supervisado (predicción de abandono) y algoritmo no supervisado (clusterización) permite a una organización no solo establecer el orden de prioridad de los clientes a contactar, sino que también es útil para determinar qué acciones pueden ser más efectivas, ya que se cuenta con un registro histórico de las ofertas realizadas a clientes de los diferentes segmentos y cuál ha sido su efecto en cuanto al abandono o retención de los mismos.

Además del beneficio que implica obtener la probabilidad de abandono y priorizar el contacto con los clientes, es importante destacar que las medidas de marketing que se realicen para retener a los clientes deseados generaran una retroalimentación de la efectividad de este proceso y permitirá conocer aún más a los consumidores.

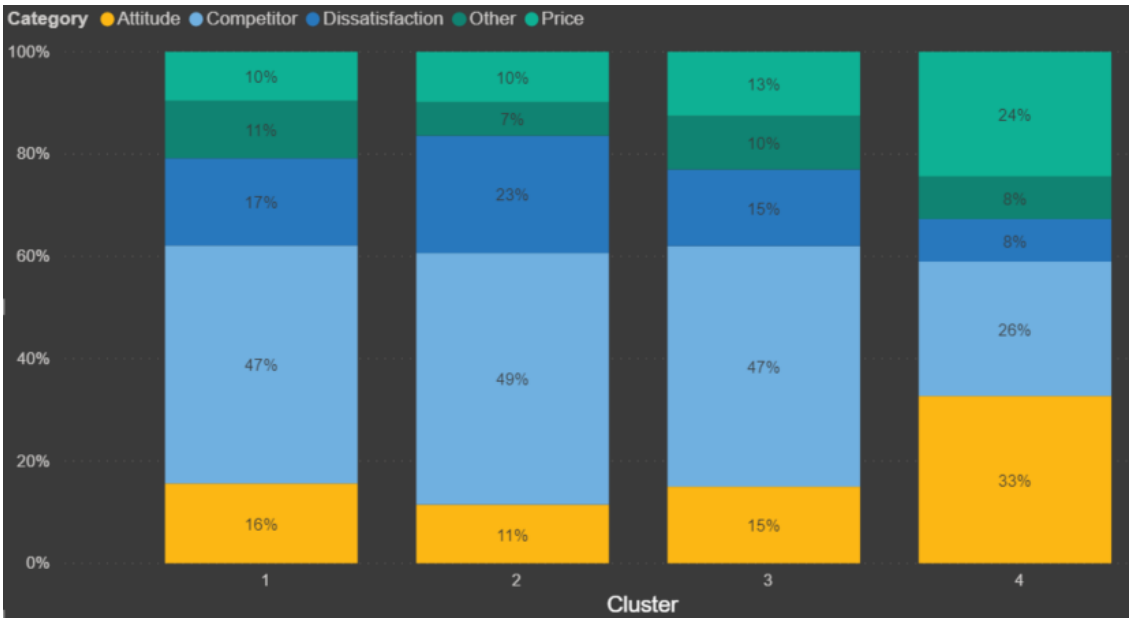
Todos estos datos, utilizados correctamente, permitirán también focalizar sobre la adquisición de clientes que se parezcan aún más a los grupos de clientes deseados por la organización. En este punto es donde se eleva el potencial de la retención y el aprendizaje que se puede realizar de los clientes actuales.

Parece que el grupo número 3 de clientes se caracteriza por tener el mayor CLTV, son las personas que más tiempo hace que están dentro de la compañía y son quienes más clientes han traído a la misma. Podría decirse que son los clientes “oro” de esta organización, según los datos de la Figura 32.

Este tipo de análisis permite decidir sobre qué clientes se hará foco en una campaña de retención. Si se observa que un cliente del grupo 3 pudiera abandonar la compañía, el personal de retención se contactará con este cliente en forma inmediata para evitar que se vaya. Debería indagarse qué tipos de ofertas o regalos puede interesarle a este conjunto de clientes, siendo una posibilidad la de proveerle un descuento a la tarifa que paga, ofrecerle un nuevo servicio gratuito por algún tiempo específico. Otra opción es que estos clientes sean los primeros en recibir algún servicio nuevo de la empresa.

También se pueden investigar las causas del abandono de los clientes por segmentos. Utilizando un gráfico de Power BI que permita ver el peso que tiene cada razón de abandono por cluster se observa lo siguiente:

Figura 33. Razones de abandono por cluster



Se evidencia en la Figura 33 que el efecto de la competencia es muy fuerte en los tres primeros clusters, pero en cluster 4 se evidencia que la actitud de la empresa, del personal de soporte o de servicio, ha sido la causa principal de enojo y, por ende, abandono de los clientes de este grupo.

Una posible acción comercial, siempre que la persona sea de valor para la empresa, sería contactar a dicho cliente y ofrecerle una disculpa por el trato recibido, asegurando que nunca volverá a ocurrir nuevamente y algún descuento a su tarifa para que vuelta a la compañía.

8.1 Lineamientos generales de acciones comerciales

Una acción comercial para retener un cliente valioso no es necesariamente una acción directa al mismo, puede implicar mejorar el tratamiento de quejas del servicio, capacitar al personal de atención al cliente, empoderar a estos empleados con mayor responsabilidad para brindarles la independencia que pudieran necesitar en algún caso, eliminar las cláusulas de castigo financiero en caso que el cliente desee dar por finalizado su contrato, crear un valor e imagen de marca sobre valores y creencias que sean respetables y aporten valor a los consumidores.

Diferentes estudios se han realizado buscando las causas y las consecuencias del abandono. Algunas cuestiones de interés que autores y autoras han encontrado son:

- “La teoría de la paradoja de la recuperación del servicio afirma que la satisfacción del cliente tras un proceso de fallo y recuperación podría ser incluso superior a la satisfacción previa al fallo“ (Varela Neira et al., 2009)
- “Una empresa cuyo objetivo sea la lealtad actitudinal de los clientes (como complemento a una lealtad comportamental) no debe sustentar su ventaja competitiva en el desarrollo de barreras al cambio negativas donde el cliente tenga que permanecer en la relación aunque no lo desee” (Varela Neira et al., 2009)
- En el caso que se produzca una relación entre barreras al cambio positivas y un cliente satisfecho, es probable que exista un efecto potenciador de esta

circunstancia, lo que repercute en una mayor probabilidad de retención del cliente (Varela Neira et al., 2009)

- "Barreras al cambio 'duras' o negativas, como un contrato extenso, pueden conllevar a un aumento de la probabilidad de cambio de proveedor por parte del cliente. Aun así, no todas las barreras al cambio positivas son igualmente útiles en todas las industrias y sectores. La calidad del servicio, la creatividad y la innovación de la empresa percibida y valorada por el cliente pueden generar el éxito de una compañía en la industria de las telecomunicaciones" (Malhotra & Kubowicz Malhotra, 2013).

8.2 Ejemplo acción comercial

Utilizando las predicciones del modelo *predi775* para ordenar los clientes con mayor probabilidad de abandono primero y los segmentos obtenidos anteriormente, podríamos tener una tabla de clientes como la que sigue:

Tabla 25. Prioridad de contacto clientes por probabilidad de abandono

CUSTOMERID	CLUSTER	PREDI775	CLTV
2656-TABEH	1	0,953651	2650
2454-RPBRZ	1	0,948282	2709
7932-WPTDS	1	0,944076	5377
2265-CYWIV	2	0,939869	4708
4795-KTRTH	2	0,929620	4287
9282-IZGQK	1	0,924388	5679
8098-LLAZX	1	0,922870	3677
0334-GDDSO	1	0,915624	2818
0576-WNXXC	1	0,914736	2999
4844-JJWUY	1	0,913407	5359
9647-ERGBE	3	0,910982	4664
1157-BQCUW	1	0,890078	2591
9605-WGJVW	1	0,888236	2422
3755-JBMNH	1	0,881599	3327
4614-NUVZD	1	0,879423	2075
5955-ERIHD	1	0,877406	2206
7273-TEFQD	1	0,866954	2314
5569-OUICF	1	0,866514	4994
7240-ETPTR	1	0,865599	4359
		TOTAL	**69215**

Se muestran los primeros 19 clientes para simplificar el ejemplo.

En este caso, con la información de la Tabla 25, si no se tuviera en cuenta el segmento (cluster) al que pertenece el cliente, en una campaña de retención se decidiría contactar a los clientes de dicha tabla. Estas personas no son necesariamente el objetivo comercial de la compañía ni aquellos que más valor proveen.

Con un enfoque diferente, se podría primero contactar aquellos clientes que sería más perjudicial perder desde una perspectiva económica. Con este enfoque, y asumiendo

que el cluster 3 es el objetivo a preservar, la lista de personas a contactar seria la siguiente:

Tabla 26. Prioridad de contacto de clientes por cluster y probabilidad de abandono

CUSTOMERID	CLUSTER	PREDI775	CLTV
9647-ERGBE	3	0,910982	4664
6646-VRFOL	3	0,833748	5860
1587-FKLZB	3	0,810534	4479
5469-CTCWN	3	0,789854	4334
5889-JTMUL	3	0,766008	5863
4803-AXVYP	3	0,700412	3474
7279-BUYWN	3	0,696520	3380
6362-QHAFM	3	0,648780	5106
4690-PKDQG	3	0,639136	4121
9835-ZIITK	3	0,628969	4281
2267-WTPYD	3	0,601438	6015
8276-MQBYC	3	0,590534	6204
3886-CERTZ	3	0,541339	4261
0397-GZBBC	3	0,522299	5390
7957-RYHQD	3	0,481857	3207
8938-UMKPI	3	0,477104	4804
2929-QNSRW	3	0,448208	4886
6686-YPGHK	3	0,417076	3645
1264-BYWMS	3	0,389943	4089
		TOTAL	**88063**

Para la creación de la Tabla 26 se ha asumido una tasa de corte de 0,3. Es decir, que si la predicción del modelo es que la probabilidad del cliente se encuentra por encima del 0,3 se considera que el mismo abandonará la empresa.

Además, si se compara el CLTV de la primera lista de clientes a contactar y la segunda, en la cual solo se incluyen clientes del cluster 3, se obtienen grandes diferencias:

- CLTV de los clientes de Tabla 25: 69215
- CLTV de los clientes de Tabla 26: 88063
- Diferencia: 18848

Es notable la diferencia de valor que se conserva si los esfuerzos de retener a 19 clientes se enfocan bajo una modalidad de complementariedad entre probabilidad de abandono y clustering. La empresa lograría que el rendimiento de los recursos enfocados en evitar el abandono de los clientes sea más alto, siempre que contacte a los clientes adecuados y a los que le es más conveniente retener.

Esta misma prioridad se puede realizar para todos los clientes que la empresa quiera contactar porque tienen alta probabilidad de abandono. Simplemente se deben ordenar por cluster y probabilidad, con la restricción de recursos financieros y de otro tipo que tenga en el momento que decida realizar estas acciones de retención.

La importancia del uso del cluster radica en que la organización utilizará mejor sus recursos, ya que desarrollar una acción comercial de retención para clientes del mismo segmento debería ser menos costosa que múltiples acciones a diferentes clientes de distintos segmentos. Además, al ser clientes que se parecen entre sí, es esperable que el mismo tipo de oferta surja efectos similares. Por lo que la construcción y la implementación de una campaña de retención a un único segmento debería ser más rentable y más precisa que acciones improvisadas a cada cliente o sin diferenciarlos por grupos homogéneos.

9 Conclusiones y trabajo futuro

Como se ha visto, el estudio de la probabilidad de abandono de clientes permite a una empresa aprender de sus errores y de sus omisiones. También, gestionándolo en forma correcta y eficiente, le permite utilizar mejor sus recursos en los clientes que desea mantener y aprender aun más de ellos.

También permite descubrir qué es lo que puede provocar el abandono de un cliente, si es por la existencia de una competencia feroz, por una mala gestión de los servicios, de las quejas, de soporte técnico, del precio de las tarifas, o por los motivos que fueran.

Lo importante de todo este estudio es que esta información puede utilizarse para mejorar la relación con el cliente y, como consecuencia, la salud de la empresa.
Esta información debe ser accesible, debe ser parte de los datos del sistema CRM para que permitan a la compañía aprovechar el potencial que tienen.

Una posible ampliación de este trabajo sería contar con más variables de consumo, de relación con el cliente como llamadas por quejas, resolución de inconvenientes y otro tipo de variables similares, para estudiar el efecto que tiene la relación cliente-empresa en el abandono de clientes con mayor precisión.

Como trabajo futuro podría estudiarse la opción de generar un proceso de probabilidad de abandono de clientes en tiempo real, con actualizaciones constantes derivadas de cada interacción con los clientes, como han propuesto algunos autores (Balle et al., s. f.).

Otra posibilidad es estudiar la elasticidad de los clientes en cuanto a variaciones a las diferentes variables más importantes que pueden generar el abandono.

También sería interesante realizar estudios tipo “A/B testing” para evidenciar el uso de diferentes ofertas o acciones de retención y con ello establecer cuáles son más efectivas según el perfil del cliente.

10 Bibliografía

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242-254. https://doi.org/10.1016/j.neucom.2016.12.009

Amin, A., Shah, B., Khattak, A. M., Lopes Moreira, F. J., Ali, G., Rocha, A., & Anwar, S. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. *International Journal of Information Management*, *46*, 304-319. https://doi.org/10.1016/j.ijinfomgt.2018.08.015

Ascarza, E. (s. f.). *Targeting high risk customers might be ineffective*. 72.

Aydin, S., & Özer, G. (2005). The analysis of antecedents of customer loyalty in the Turkish mobile telecommunication market. *European Journal of Marketing*, *39*(7/8), 910-925. https://doi.org/10.1108/03090560510601833

Balle, B., Casas, B., Catarineu, A., Gavalda, R., & Manzano-Macho, D. (s. f.). *The Architecture of a Churn Prediction System Based on Stream Mining*. 13.

Bolancé, C., Montserrat, G., & Padilla-Barreto, A. E. (2016). Predicting Probability of Customer Churn in Insurance. En *Modeling and Simulation in Engineering, Economics and Management* (Vol. 254, pp. 82-91). Springer International Publishing. https://doi.org/10.1007/978-3-319-40506-3_9

Calviño, A. (2019). *Materiales Asignatura "Técnicas y metodología de la minería de datos (SEMMA)"—Máster Minería de Datos e Inteligencia Negocio (UCM).*

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, *61*, 304-323. https://doi.org/10.1016/j.iref.2018.03.008

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27-36. https://doi.org/10.1016/j.dss.2016.11.007

Devriendt, F., Berrevoets, J., & Verbeke, W. (2019). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, S0020025519312022. https://doi.org/10.1016/j.ins.2019.12.075

Gallo, A. (s. f.). *The Value of Keeping the Right Customers*. 3.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, *197*(1), 402-411. https://doi.org/10.1016/j.ejor.2008.06.027

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research*, *8*(6), p224. https://doi.org/10.5539/ibr.v8n6p224

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425. https://doi.org/10.1016/j.eswa.2011.08.024

IE Catedra de Fidelización, Gonçalves, D., & InLoyalty. (2018). *Barómetro Fidelización 2018*.

Jones, T. O., & Sasser, W. E. (s. f.). *Why Satisfied MARKETRESEARCH Customers Defect*. 25.

Kim, H.-S., & Yoon, C.-H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, *28*(9-10), 751-765. https://doi.org/10.1016/j.telpol.2004.05.013

Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, *43*(2), 276-286. https://doi.org/10.1509/jmkr.43.2.276

Malhotra, A., & Kubowicz Malhotra, C. (2013). Exploring switching behavior of US mobile service customers. *Journal of Services Marketing*, *27*(1), 13-24. https://doi.org/10.1108/08876041311296347

McIlroy, A., & Barnett, S. (2000). Building customer relationships: Do discount cards work? *Managing Service Quality: An International Journal*, *10*(6), 347-355. https://doi.org/10.1108/09604520010351491

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection Measuring and Undertanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *Vol. XLIII*, 204–211.

Núñez, D. L. (2015). *Modelos predictivos del churn – abandono de clientes – para operadores de telecomunicaciones*. 68.

Portela, J. (2020). *Materiales Asignatura "Técnicas de Machine Learning"—Máster Minería de Datos e Inteligencia Negocio (UCM).*

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A Proposed Churn Prediction Model. *International Journal of Engineering*, *2*(4), 5.

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. *Journal of Service Research*, *19*(2), 123-141. https://doi.org/10.1177/1094670515616376

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258-1268. https://doi.org/10.1016/j.indmarman.2014.06.016

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423. https://doi.org/10.1111/1467-9868.00293

V. Kumar. (2018). *Customer relationship management: Concept, strategy, and tools*. Springer Berlin Heidelberg.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1-9. https://doi.org/10.1016/j.simpat.2015.03.003

Varela Neira, C., Vázquez Casielles, R., & Iglesias Argüelles, V. (2009). Comportamiento de abandono de la relación de un cliente con la empresa en un contexto de fallo y recuperación del servicio. *Cuadernos de Economía y Dirección de la Empresa*, *12*(40), 143-169. https://doi.org/10.1016/S1138-5758(09)70045-3

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211-229. https://doi.org/10.1016/j.ejor.2011.09.031

Zhu, B., Baesens, B., & vanden Broucke, S. K. L. M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, *408*, 84-99. https://doi.org/10.1016/j.ins.2017.04.015

11 Anexo I – Resultados de selección de variables

1) Variables seleccionadas por el proceso en R con el método “stepwise”:

- Age
- Contract_Month_to_Month
- Contract_One_Year
- Device_Protection_Plan
- Internet_Type_Cable
- Internet_Type_DSL
- Internet_Type_Fiber_Optic
- Married
- Monthly_Charge
- Multiple_Lines
- Number_of_Dependents_0
- Number_of_Referrals_0
- Number_of_Referrals_1
- Number_of_Referrals_2
- Number_of_Referrals_3
- Number_of_Referrals_4
- Number_of_Referrals_5
- Number_of_Referrals_6
- Number_of_Referrals_7
- Number_of_Referrals_8
- Number_of_Referrals_9
- Offer_Offer_A
- Offer_Offer_D
- Online_Backup
- Online_Security
- Paperless_Billing
- Payment_Method_Bank_Withdrawal
- Payment_Method_Credit_Card
- Phone_Service
- Population
- Premium_Tech_Support
- Streaming_Movies
- Streaming_Music
- Streaming_TV
- Tenure_in_Months
- Total_Charges
- Total_Refunds

2) En SAS, utilizando Proc logistic con selection=stepwise

- Monthly_Charge
- Population
- Tenure_in_Months
- Total_Charges
- Contract
- Device_Protection_Plan
- Internet_Type
- Married
- Multiple_Lines
- Number_of_Dependent
- Number_of_Referrals
- Offer
- Online_Backup
- Online_Security
- Paperless_Billing
- Payment_Method
- Phone_Service
- Premium_Tech_Suppor
- Streaming_Movies
- Streaming_Music
- Streaming_TV

3) En SAS, utilizando Proc logistic con selection=score (mejor selección):

- Age
- Monthly_Charge
- Population
- Tenure_in_Months
- Total_Charges
- Total_Refunds
- Online_Backup
- Online_Security
- Paperless_Billing
- Phone_Service
- Premium_Tech_Support

4) En SAS, con la macro %randomselectlog (Portela, 2020) con la mayor cantidad de apariciones:

- Contract
- Number_of_Dependent
- Number_of_Referrals
- Monthly_Charge
- Population
- Tenure_in_Months
- Total_Charges
- Total_Refunds
- Online_Backup
- Online_Security
- Paperless_Billing
- Phone_Service
- Premium_Tech_Suppor
- Streaming_TV

En los 4 métodos de selección, uniendo el utilizado en R y los 3 correspondientes a SAS, las siguientes variables son escogidas:

- Monthly_Charge
- Online_Backup
- Online_Security
- Paperless_Billing
- Phone_Service
- Population
- Premium_Tech_Support
- Tenure_in_Months
- Total_Charges

12 Anexo II – Configuración de modelos de predicción de abandono

En este anexo se presentan los procesos de configuración de los hiperparámetros de los algoritmos utilizados para la predicción de abandono de clientes.

12.1 Redes neuronales

En R se ha utilizado la grilla de la librería caret para combinar diferentes valores de hiperparámetros y observar el comportamiento de estos con el conjunto de datos.

Con el uso del *grid,* se han construido redes neuronales diferentes redes, con un número de 3, 5, 7, 9, 11 y 13 nodos, en combinación con *learning rates* de 0.1, 0.01 y 0.001. Se han realizado un total de 100 iteraciones con el método “avNNet”.

Los resultados se exhiben en la Tabla 27:

Tabla 27. Resultados de grilla de hiperparámetros de redes neuronales (R)

SIZE	DECAY	ACCURACY	KAPPA
3	0.001	0.8428801	0.5683469
3	0.010	0.8444139	0.5826121
3	0.100	0.8466287	0.5929812
5	0.001	0.8404100	0.5614460
5	0.010	0.8435331	0.5834008
5	0.100	0.8448111	0.5855536
7	0.001	0.8402390	0.5636807
7	0.010	0.8440445	0.5840168
7	0.100	0.8442715	0.5836278
9	0.001	0.8410625	0.5722964
9	0.010	0.8408634	0.5782856
9	0.100	0.8420566	0.5779697
11	0.001	0.8394153	0.5644332
11	0.010	0.8417159	0.5802629
11	0.100	0.8410622	0.5778528
13	0.001	0.8376829	0.5575673
13	0.010	0.8396138	0.5737217
13	0.100	0.8393305	0.5712114

La red con 3 nodos y el *decay* igual a 0.1 parece ser la mejor opción para estos datos observando únicamente el accuracy.

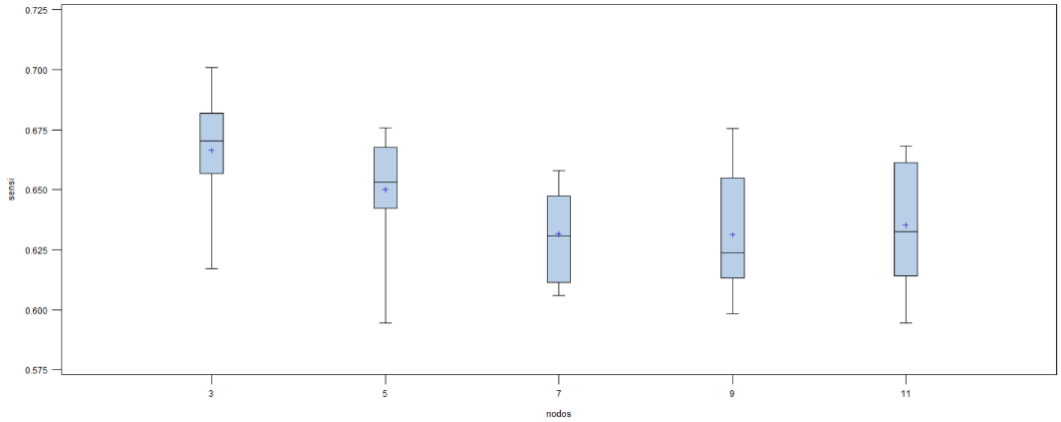
En SAS para la creación de redes neuronales se utiliza como primera aproximación la macro *%variar* en conjunto con la macro *%neuralbinariabasica* (Portela, 2020), para realizar pruebas de variación de nodos. Este proceso se realiza una vez utilizando el algoritmo de Levenberg-Marquardt (Levmar) y, por otro lado, algunas pruebas adicionales con Backpropagation y variaciones en los hiperparámetros de *momentum* y *learning rate*.

12.1.1 Levenberg-Marquardt

Resultados obtenidos con la variación del número de nodos, utilizando 3, 5, 7, 9 y 11 nodos:

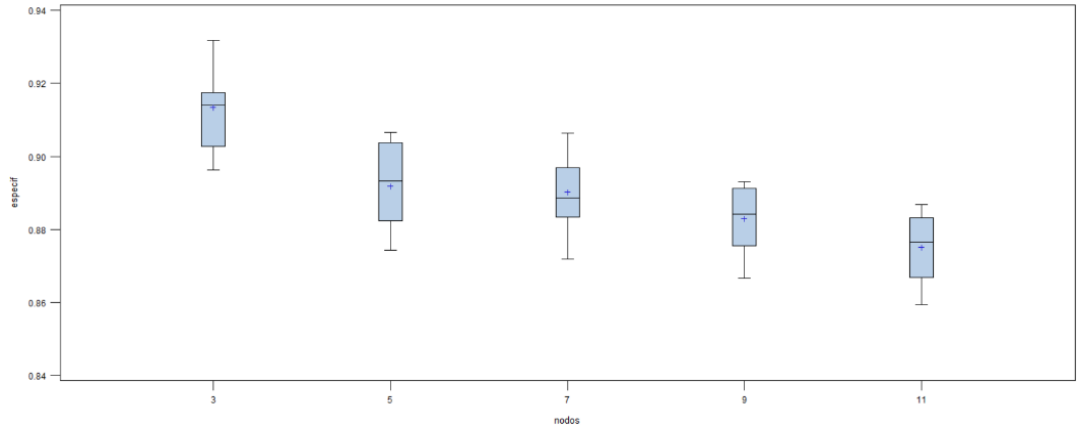
- Sensibilidad:

Figura 34.Sensibilidad de redes neuronales con Levenberg-Marquardt (SAS)



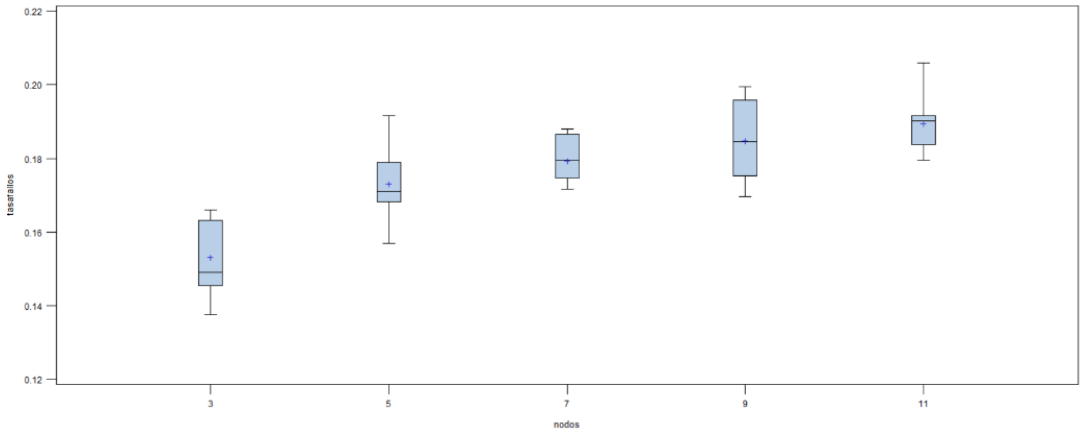
- Especificidad

Figura 35. Especificidad de redes neuronales con Levenberg-Marquardt (SAS)



- Tasa de fallos:

Figura 36.Tasa de fallos de redes neuronales con Levenberg-Marquardt (SAS)



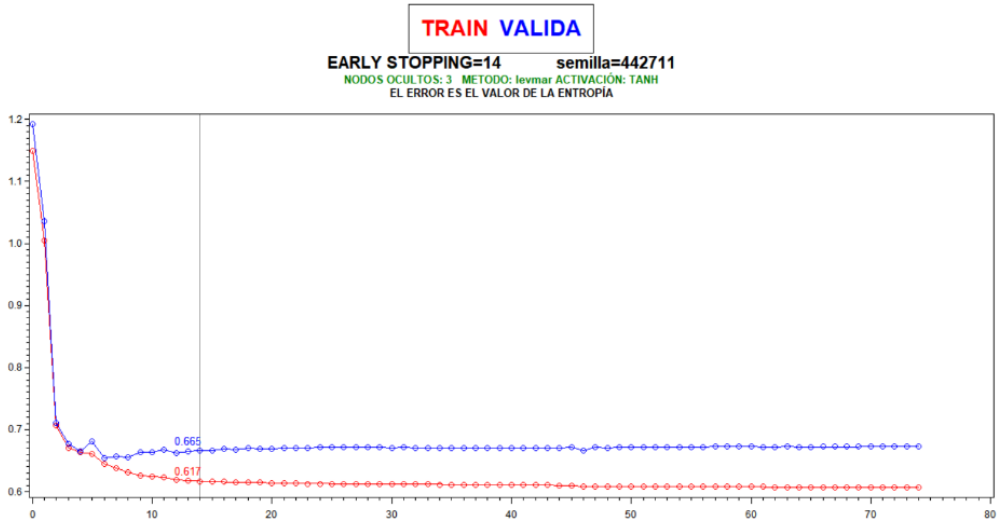
Es evidente que el uso de menor número de nodos, para este caso, funciona mejor, según los resultados de las FigurasFigura **34**, Figura **35** y Figura **36**.

Se realiza un análisis de early stopping con la macro *%redneuronalbinaria* para las redes compuestas por 3, 5 y 7 nodos.

Para los 3 casos se han utilizado 4 semillas diferentes y se obtuvieron resultados similares. Los gráficos que se exponen a continuación corresponden a una de dichas pruebas:

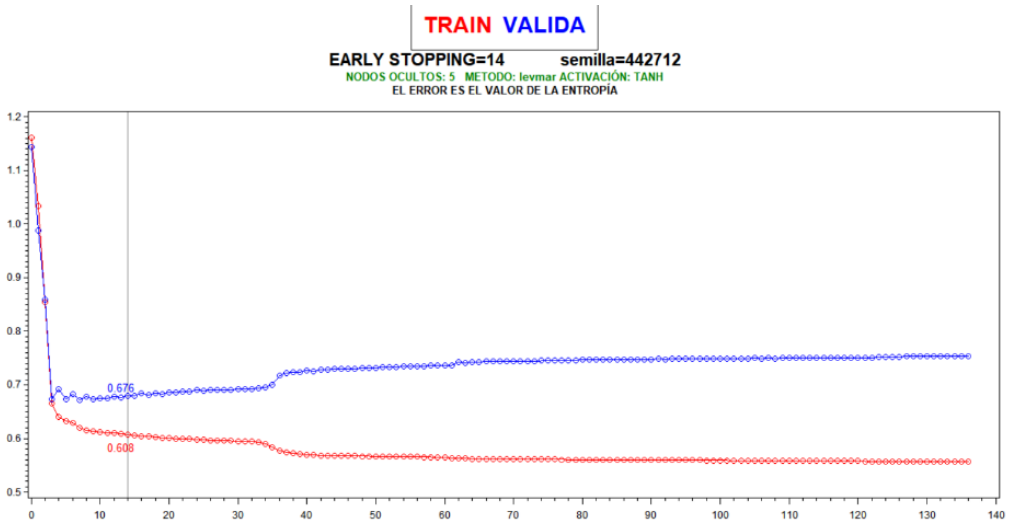
- 3 nodos

Figura 37. Estudio early stopping red de 3 nodos



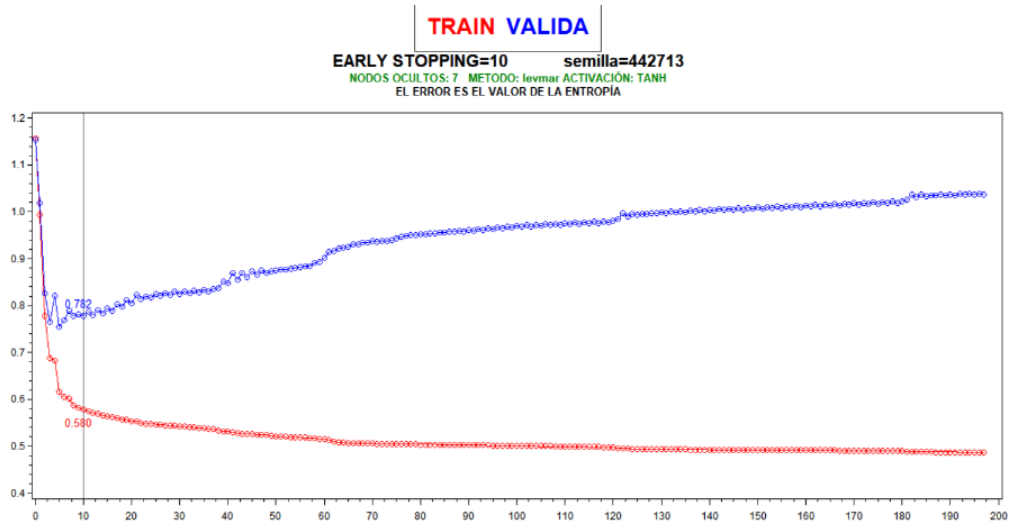
- 5 nodos:

Figura 38. Estudio early stopping red de 5 nodos



- 7 nodos:

Figura 39. Estudio early stopping red de 7 nodos



En los 3 casos con un early stopping menor a 15 parece suficiente, según los resultados visibles de las figuras Figura **37**, Figura **38** y Figura **39**.

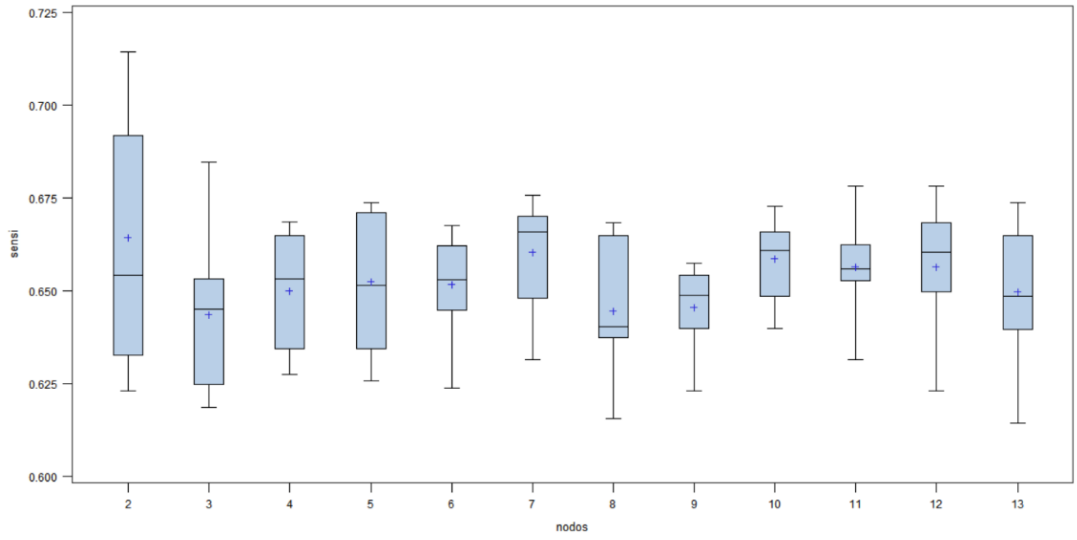
12.1.2 Backpropagation

Se utiliza la macro *%variar* con *%neuralbinariabasica,* con una variación de nodos de 2 a 13 de uno en uno. Se han realizado diferentes pruebas en combinaciones de *momentum* y *learning rates* diferentes.

A continuación, se exhiben algunos gráficos correspondientes a la mejor configuración elaborada con este algoritmo: momentum = 0,3 y learning rate = 0,2:

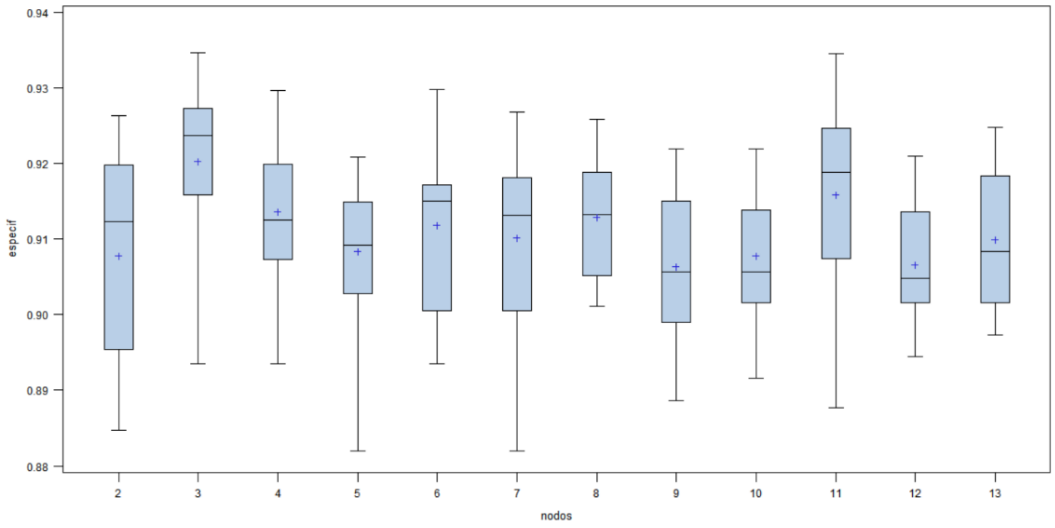
- Sensibilidad:

Figura 40.Sensibilidad de redes neuronales con Backpropagation (SAS)



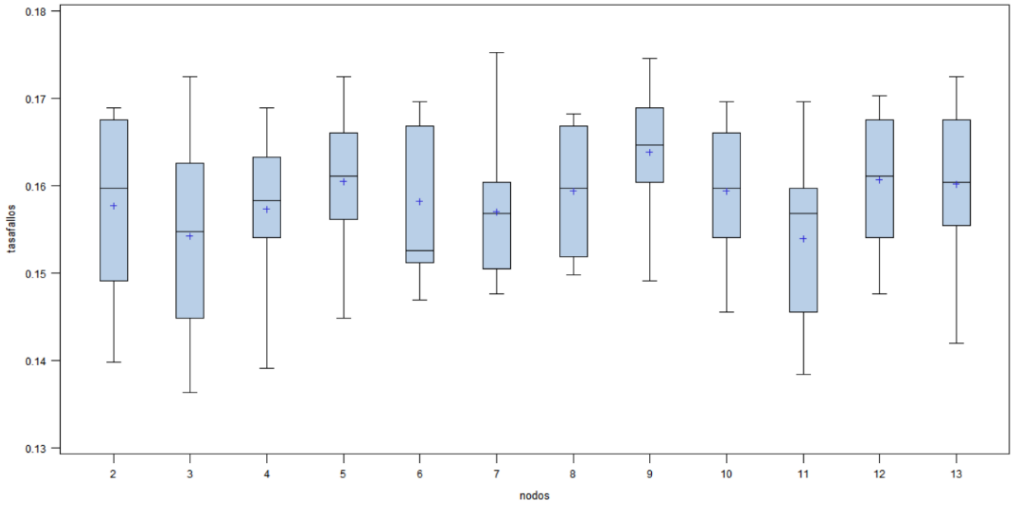
- Especificidad:

Figura 41. Especificidad de redes neuronales con Backpropagation (SAS)



- Tasa de fallos:

Figura 42. Tasa de fallos de redes neuronales con Backpropagation (SAS)



Parecería que entre 4 y 7 nodos esta red funciona bien y que no es necesario aumentar el número de nodos, observando los diagramas de cajas de las Figuras Figura **40**, Figura **41** y Figura **42**.

12.2 Random forest

En Rstudio, se utiliza la grilla con la librería *caret* para realizar un acercamiento al posible mejor modelo de *random forest* aplicable a la predicción de abandono de cliente.

Se utilizan los siguientes valores para el parámetro *mtry*: 3, 9, 15, 21, 25, 31, 37, 43, 49, 52 y 56.
El valor 56 considera el caso de *bagging*, ya que se utilizaría la totalidad de las variables del conjunto de datos (son 56 variables debido a que las categorías de las variables nominales ahora son variables “dummy”).

Como primera prueba de este algoritmo, se utiliza un total de 1000 árboles (*ntree*), un tamaño de muestra de 200 (*sampsize*) y un tamaño de nodo de 10 (*nodesize*).

Los valores obtenidos para los diferentes modelos de *random forest* y *bagging* son:

Tabla 28. Resultados de 1° grilla de hiperparámetros de random forest (R)

MTRY	ACCURACY	KAPPA
3	0.8164157	0.4417351
9	0.8296217	0.5044900
15	0.8357255	0.5325029
21	0.8391335	0.5437561
25	0.8391331	0.5462894
31	0.8405538	0.5528144
37	0.8414061	0.5548821
43	0.8441030	0.5643173
49	**0.8449556**	**0.5684611**
52	0.8435358	0.5671104
56	0.8435352	0.5661026

Con los resultados de la Tabla 28, parece que el uso de 49 variables para este caso es el mejor tomando en cuenta la exactitud de las predicciones.

Se realiza una prueba similar, variando el número de árboles a crear (*ntree*), siendo el nuevo valor igual a 300 y se obtiene el siguiente resultado:

Tabla 29. Resultados de 2° grilla de hiperparámetros de random forest (R)

MTRY	ACCURACY	KAPPA
3	0.8406918	0.5568037
9	0.8428216	0.5694355
15	0.8440999	0.5774163
21	0.8429638	0.5754340
25	**0.8462292**	**0.5851419**
31	0.8448093	0.5820645
37	0.8426795	0.5772316
43	0.8418277	0.5746337
49	0.8429638	0.5790140
52	0.8399824	0.5703403
56	0.8415440	0.5760210

En este caso de la Tabla 29, el mejor número de variables es 25.

Utilizando la información de estos modelos, se revisa la importancia de cada variable en la Tabla 30:

Tabla 30. Importancia de variables random forest (R)

VARIABLES	MEANDECREASEACCURACY	MEANDECREASEGINI
NUMBER_OF_REFERRALS_1	55,4884461	103,285165
AGE	44,1270355	189,163257
NUMBER_OF_DEPENDENTS_0	36,4976246	82,7268511
CONTRACT_MONTH_TO_MONTH	35,2527404	360,434684
MONTHLY_CHARGE	34,1622457	180,065818
TOTAL_CHARGES	29,5716879	125,316333
TENURE_IN_MONTHS	28,4517091	181,820684
TOTAL_LONG_DISTANCE_CHARGES	24,4844273	105,550313
AVG_MONTHLY_GB_DOWNLOAD	17,7439678	84,1700663

PAYMENT_METHOD_CREDIT_CARD	15,9135621	31,9797678
INTERNET_TYPE_FIBER_OPTIC	14,9326927	65,1039299
AVG_MONTHLY_LONG_DISTANCE_CHARGE	14,8031936	86,7556099
CONTRACT_TWO_YEAR	14,4586203	89,7455852
NUMBER_OF_REFERRALS_0	14,2673851	17,4748587
NUMBER_OF_REFERRALS_9	13,8550746	4,6606312
PREMIUM_TECH_SUPPORT	13,6163904	12,7815517
ONLINE_SECURITY	12,3447624	15,9729244
STREAMING_MUSIC	11,0684282	16,3165252
MARRIED	10,3932545	9,0420343
NUMBER_OF_REFERRALS_8	10,2512781	3,8984011
PARTNER	9,5564149	9,3846597
UNLIMITED_DATA	8,9093205	8,6559044
STREAMING_TV	8,7189192	9,9275128
NUMBER_OF_DEPENDENTS_1	8,1099337	5,0494049
INTERNET_TYPE_DSL	8,0627861	5,6794383
ONLINE_BACKUP	7,6625754	8,3003973
INTERNET_TYPE_NONE	7,5722728	11,7326787
CLTV	7,3624265	99,0932795
OFFER_OFFER_B	7,1061274	6,65038
CONTRACT_ONE_YEAR	7,0597334	35,4867342
PAPERLESS_BILLING	6,9524048	13,9639785
PAYMENT_METHOD_MAILED_CHECK	6,6599821	11,1612877
NUMBER_OF_REFERRALS_7	6,6470933	2,304207
STREAMING_MOVIES	6,5381026	8,2892499
PAYMENT_METHOD_BANK_WITHDRAWAL	6,264542	8,441709
MULTIPLE_LINES	5,9721496	7,9416948
DEVICE_PROTECTION_PLAN	5,4434902	6,2070834
INTERNET_TYPE_CABLE	5,3465295	6,2160695
NUMBER_OF_REFERRALS_6	5,2472534	2,2426698
POPULATION	5,1874336	109,784734
NUMBER_OF_DEPENDENTS_2	4,7179555	3,0004514
NUMBER_OF_REFERRALS_10	4,2794752	9,6542E-08
NUMBER_OF_DEPENDENTS_3	3,6670301	3,2032576
OFFER_OFFER_E	2,9950085	7,0937727
TOTAL_EXTRA_DATA_CHARGES	2,9674502	21,8459943
NUMBER_OF_REFERRALS_4	2,0578487	2,6905103
PHONE_SERVICE	2,0342845	2,0706425
NUMBER_OF_REFERRALS_2	1,980957	3,0597353
OFFER_NONE	1,7957215	7,0908431
GENDER_FEMALE	-3,4107E-08	7,6223138
OFFER_OFFER_A	-5,2829E-08	1,5001106
TOTAL_REFUNDS	-5,4802E-08	11,1900196
NUMBER_OF_REFERRALS_3	-8,3106E-08	2,8403009
NUMBER_OF_REFERRALS_5	-1,4312661	3,0317067
OFFER_OFFER_D	-2,3532111	3,0910529
OFFER_OFFER_C	-3,380366	3,2853728

Se han realizado pruebas con random forest similares con las primeras 20 variables de la Tabla 30.

Con el uso del *grid,* variando el hiperparámetro de *nodesize* con los valores 10, 20 y 30 se crean 3 random forest.

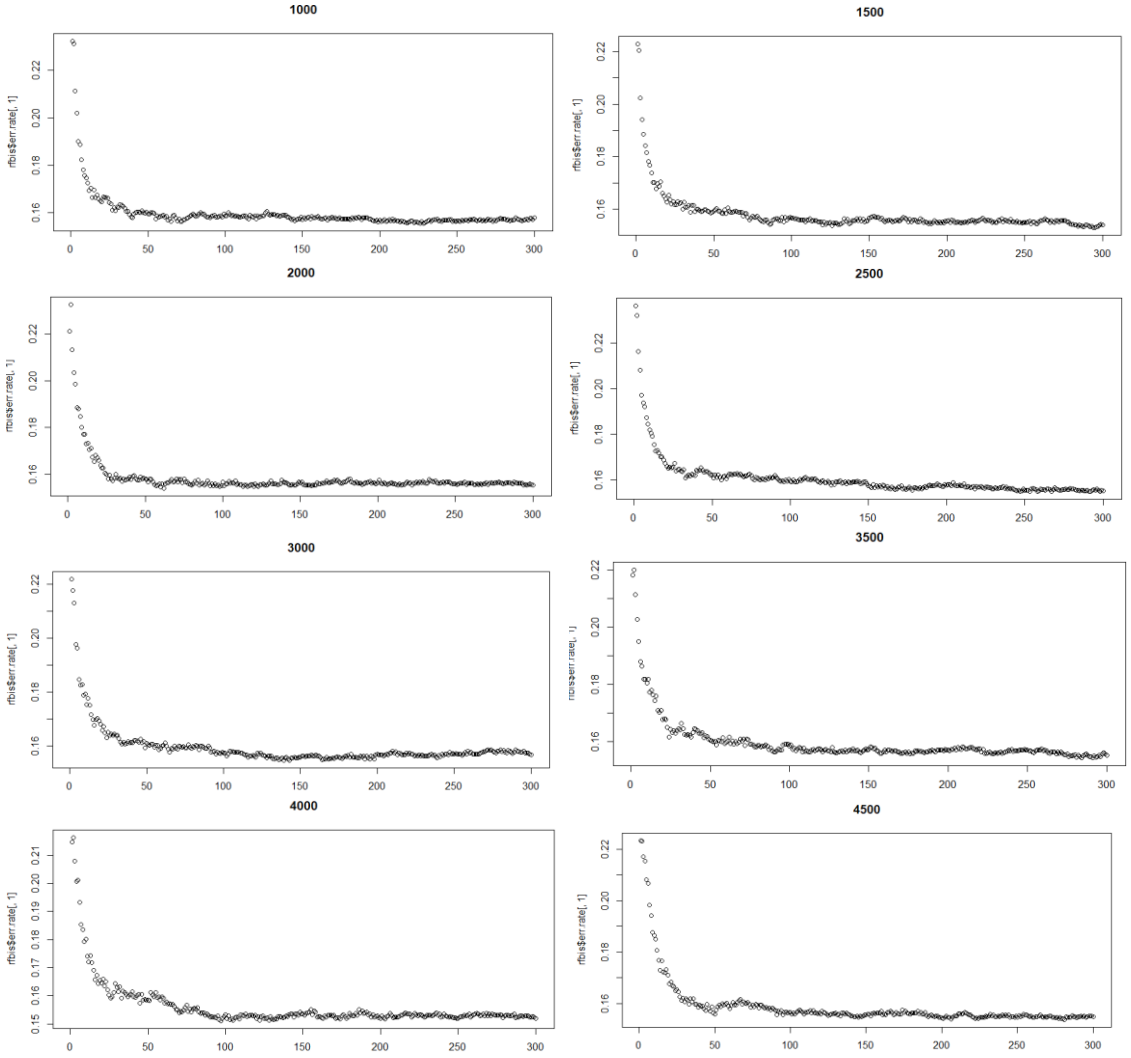
Todos han obtenido resultados de 0,84 de Accuracy, es decir, similares a los casos en los cuales se utilizan todas las variables del conjunto de datos.

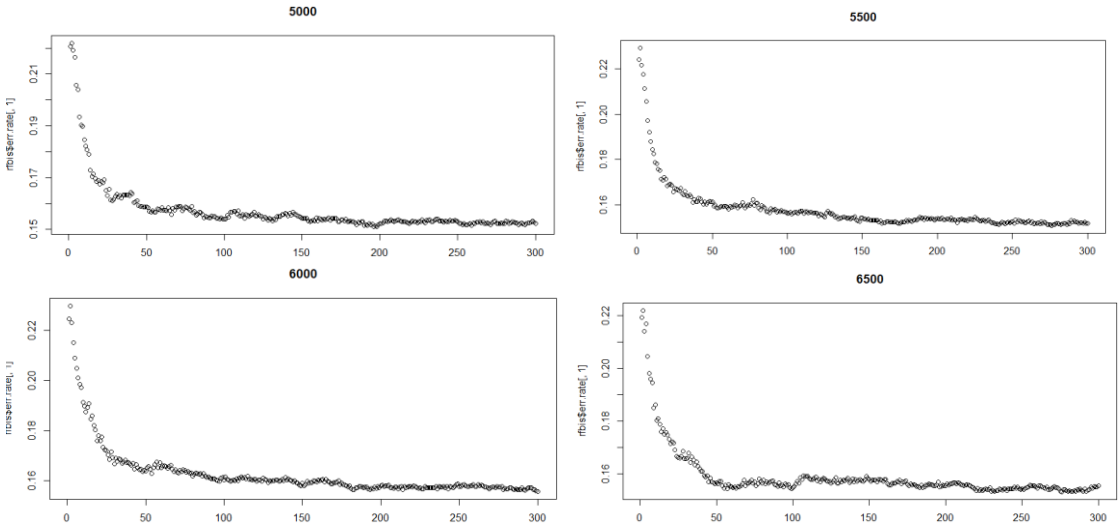
La diferencia es que variando el número de observaciones por nodo la cantidad necesaria de variables a escoger (mtry) por algoritmo varia, siendo esta de 3, 15 y 7 respectivamente para obtener el mayor valor de accuracy de dichas pruebas.

Aun variando el número de variables a escoger en cada caso, el accuracy obtenido es de 0,84 en el promedio de los modelos.

También se ha realizado un estudio del mejor número de tamaño de muestra para la creación de los diferentes conjuntos de árboles. Utilizando 15 variables a escoger para la apertura de cada nodo (*mtry*), un total de 300 árboles (*ntree*) y un tamaño de nodo de 10 (*nodesize*), se crea una secuencia que varía el número de muestra de 1000 a 6500 de 500 en 500.

Figura 43. Estudio de tamaño de muestra de random forest (R)





Observando los gráficos de la Figura 43 parece que con un tamaño de muestra de 2000 o menos inclusive, el algoritmo funciona bien. También se ha detectado que con el uso de 250 árboles no cambia demasiado el error de estimación.

Con el objetivo de estudiar la posibilidad de que el algoritmo *random forest* en SAS funcione bien para este tipo de datos, se realiza pruebas progresivas y modificaciones a los modelos probados para encontrar uno que sea útil.

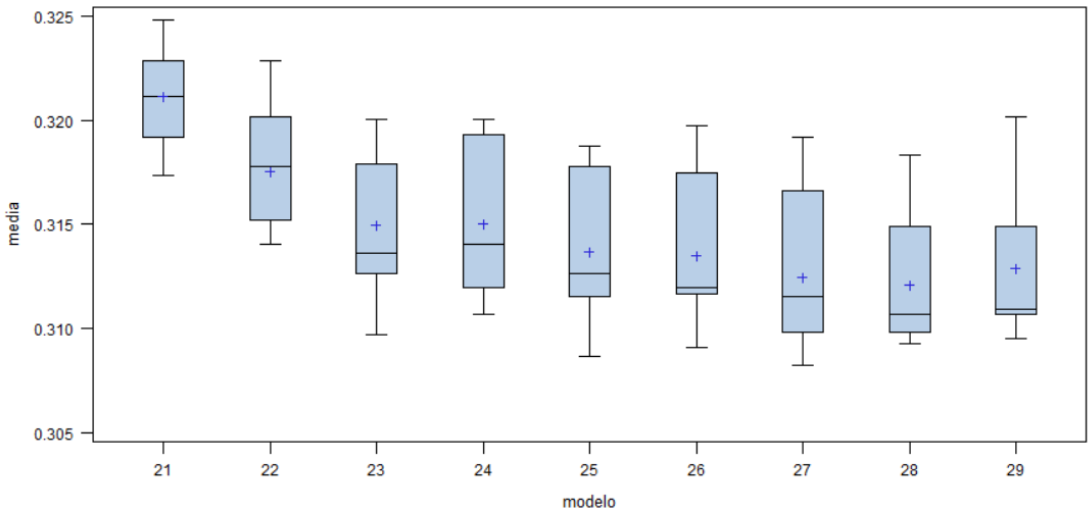
Utilizando la macro *%cruzadarandomforestbin* se generan varios modelos diferentes, con la configuración establecida en la Tabla 31.

Tabla 31. Configuración de hiperparámetros de random forest (SAS)

MODELOS	21	22	23	24	25	26	27	28	29
MAXTREES	200	200	200	300	400	400	200	150	200
VARIABLES	30	25	15	15	15	10	10	8	8
TAMHOJA	30	20	25	25	20	20	30	25	20
MAXDEPTH	20	15	10	8	8	8	10	8	6
PVALOR	0.1	0.1	0.1	0.05	0.01	0.01	0.01	0.01	0.01
PORCENBAG	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
MAXBRANCH	4	4	4	4	4	4	4	4	4

Los resultados obtenidos se grafican a continuación:

Figura 44. Tasa de fallos de random forest (SAS)



A pesar de haber realizado variaciones, no se obtiene un modelo con *random forest* suficientemente bueno como para competir al desempeño de la regresión logística y/o red neuronal, según los box-plots de la Figura 44.

A raíz que sus resultados se encuentran muy distantes al resto de los modelos generados, no se utilizará dentro del trabajo el algoritmo de *random forest* de SAS.

12.3 Gradient Boosting

En R se hace uso la grilla de combinaciones de hiperparámetros de la librería *caret*, utilizando validación cruzada, como en todos los casos anteriores y posteriores dentro del presente trabajo.

A continuación, se presentan las combinaciones realizadas:

- Shirnkage (regularización): 0.1, 0.05, 0.03, 0.01 y 0.001
- Observaciones mínimas por nodo: 5, 10, 20 y 30
- Número de árboles: 100, 500, 1000 y 5000

El resultado obtenido para el conjunto de datos es el siguiente:

Tabla 32. Resultados de 1° grilla de gradient boosting (R)

SHRINKAGE	N.MINOBSINNODE	N.TREES	ACCURACY	KAPPA
0.001	5	100	0.7346302	0.0000000
0.001	5	500	0.7346302	0.0000000
0.001	5	1000	0.7684217	0.1941452
0.001	5	5000	0.8396984	0.5498186
0.001	10	100	0.7346302	0.0000000
0.001	10	500	0.7346302	0.0000000
0.001	10	1000	0.7684217	0.1941452
0.001	10	5000	0.8396984	0.5498186
0.001	20	100	0.7346302	0.0000000
0.001	20	500	0.7346302	0.0000000
0.001	20	1000	0.7684217	0.1941452

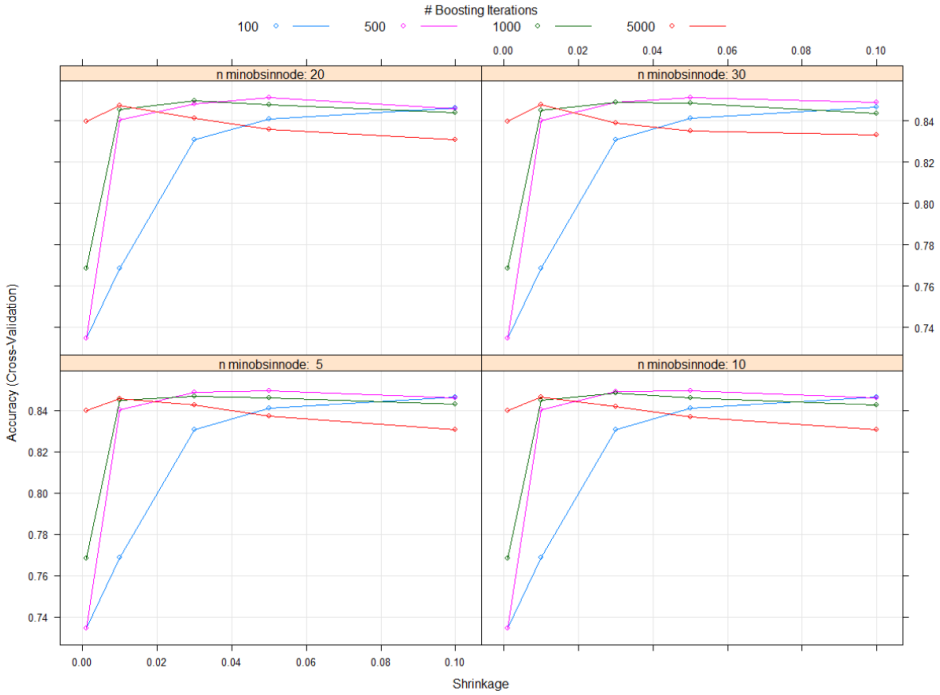
0.001	20	5000	0.8396984	0.5498186
0.001	30	100	0.7346302	0.0000000
0.001	30	500	0.7346302	0.0000000
0.001	30	1000	0.7684217	0.1941452
0.001	30	5000	0.8396984	0.5498186
0.010	5	100	0.7685637	0.1948362
0.010	5	500	0.8402666	0.5517260
0.010	5	1000	0.8446694	0.5728678
0.010	5	5000	0.8455198	0.5867888
0.010	10	100	0.7685637	0.1948362
0.010	10	500	0.8402666	0.5517260
0.010	10	1000	0.8448113	0.5731796
0.010	10	5000	0.8463719	0.5887780
0.010	20	100	0.7685637	0.1948362
0.010	20	500	0.8402666	0.5517260
0.010	20	1000	0.8455214	0.5750514
0.010	20	5000	0.8473669	0.5910978
0.010	30	100	0.7685637	0.1948362
0.010	30	500	0.8399828	0.5507623
0.010	30	1000	0.8452370	0.5743073
0.010	30	5000	0.8479339	0.5937883
0.030	5	100	0.8307543	0.5074950
0.030	5	500	0.8485022	0.5887260
0.030	5	1000	0.8466563	0.5885267
0.030	5	5000	0.8425382	0.5813889
0.030	10	100	0.8307543	0.5074950
0.030	10	500	0.8490699	0.5902621
0.030	10	1000	0.8483602	0.5932094
0.030	10	5000	0.8418281	0.5790602
0.030	20	100	0.8307543	0.5074950
0.030	20	500	0.8482182	0.5874930
0.030	20	1000	0.8497802	0.5965757
0.030	20	5000	0.8414018	0.5774894
0.030	30	100	0.8307543	0.5074950
0.030	30	500	0.8490704	0.5900919
0.030	30	1000	0.8487864	0.5944450
0.030	30	5000	0.8389883	0.5729009
0.050	5	100	0.8409768	0.5543066
0.050	5	500	0.8493536	0.5945748
0.050	5	1000	0.8459458	0.5885955
0.050	5	5000	0.8370008	0.5692655
0.050	10	100	0.8409768	0.5543066
0.050	10	500	0.8494956	0.5951265
0.050	10	1000	0.8458041	0.5879926
0.050	10	5000	0.8365756	0.5683394
0.050	20	100	0.8409768	0.5543066
0.050	20	500	0.8510574	0.5995425
0.050	20	1000	0.8477924	0.5926063

0.050	20	5000	0.8360069	0.5662190
0.050	30	100	0.8411187	0.5547881
0.050	**30**	**500**	**0.8510578**	**0.5997108**
0.050	30	1000	0.8483610	0.5947162
0.050	30	5000	0.8350126	0.5646716
0.100	5	100	0.8462310	0.5783305
0.100	5	500	0.8458039	0.5876746
0.100	5	1000	0.8429649	0.5815980
0.100	5	5000	0.8306122	0.5541367
0.100	10	100	0.8462310	0.5783305
0.100	10	500	0.8458039	0.5877118
0.100	10	1000	0.8426794	0.5808864
0.100	10	5000	0.8304705	0.5550109
0.100	20	100	0.8460888	0.5779010
0.100	20	500	0.8459463	0.5877251
0.100	20	1000	0.8438167	0.5832325
0.100	20	5000	0.8308968	0.5543454
0.100	30	100	0.8466566	0.5798979
0.100	30	500	0.8490696	0.5971479
0.100	30	1000	0.8433892	0.5831168
0.100	30	5000	0.8330261	0.5611674

Para las combinaciones de la Tabla 32, el mejor caso del algoritmo *Gradient Boosting* se obtiene con el uso de 500 árboles, un *shrinkage* de 0,05 y un número de observaciones de 30 por cada nodo. Con estas combinaciones, se ha obtenido un accuracy de 0,8510.

Si se observa gráficamente las combinaciones utilizadas se podría ajustar algún hiperparámetros para obtener un *Gradient Boosting* más preciso:

Figura 45. Gráfico de resultados de 1° grilla de gradient boosting (R)



Como puede detectarse en la Figura 45, se alcanza un máximo de accuracy en los valores del hiperparámetro de regularización que inician en 0,02 hasta el 0,06.

Con esta nueva información se realizará un nuevo proceso de combinación de algoritmos utilizando la grilla, pero con los siguientes *shrinkages*: 0.02, 0.03, 0.04, 0.05 y 0.06.
Los resultados de esta modificación se presentan a continuación:

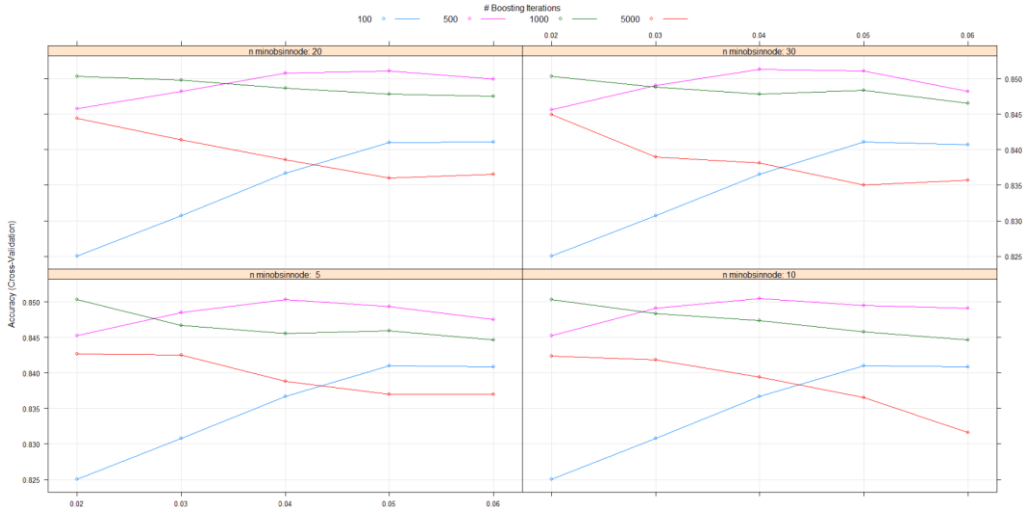
Tabla 33. Resultados de 2° grilla de gradient boosting (R)

SHRINKAGE	N.MINOBSINNODE	N.TREES	ACCURACY	KAPPA
0.02	5	100	0.8250751	0.4747160
0.02	5	500	0.8452371	0.5744485
0.02	5	1000	0.8503475	0.5958493
0.02	5	5000	0.8426800	0.5811793
0.02	10	100	0.8250751	0.4747160
0.02	10	500	0.8452371	0.5744485
0.02	10	1000	0.8503477	0.5962832
0.02	10	5000	0.8423955	0.5802827
0.02	20	100	0.8250751	0.4747160
0.02	20	500	0.8458049	0.5761735
0.02	20	1000	0.8503475	0.5962958
0.02	20	5000	0.8443841	0.5853868
0.02	30	100	0.8250751	0.4747160
0.02	30	500	0.8456629	0.5755673
0.02	30	1000	0.8503481	0.5968499
0.02	30	5000	0.8449517	0.5870275
0.03	5	100	0.8307543	0.5074950
0.03	5	500	0.8485022	0.5887260
0.03	5	1000	0.8466563	0.5885267
0.03	5	5000	0.8425382	0.5813889
0.03	10	100	0.8307543	0.5074950
0.03	10	500	0.8490699	0.5902621
0.03	10	1000	0.8483602	0.5932094
0.03	10	5000	0.8418281	0.5790602
0.03	20	100	0.8307543	0.5074950
0.03	20	500	0.8482182	0.5874930
0.03	20	1000	0.8497802	0.5965757
0.03	20	5000	0.8414018	0.5774894
0.03	30	100	0.8307543	0.5074950
0.03	30	500	0.8490704	0.5900919
0.03	30	1000	0.8487864	0.5944450
0.03	30	5000	0.8389883	0.5729009
0.04	5	100	0.8367172	0.5389841
0.04	5	500	0.8503480	0.5964172
0.04	5	1000	0.8455199	0.5862134
0.04	5	5000	0.8388473	0.5715068
0.04	10	100	0.8367172	0.5389841
0.04	10	500	0.8504895	0.5967603

0.04	10	1000	0.8473656	0.5913497
0.04	10	5000	0.8394148	0.5737889
0.04	20	100	0.8367172	0.5389841
0.04	20	500	0.8507734	0.5976503
0.04	20	1000	0.8486445	0.5947043
0.04	20	5000	0.8385628	0.5716862
0.04	30	100	0.8365753	0.5384969
0.04	**30**	**500**	**0.8513419**	**0.5996028**
0.04	30	1000	0.8477923	0.5927556
0.04	30	5000	0.8381363	0.5711055
0.05	5	100	0.8409768	0.5543066
0.05	5	500	0.8493536	0.5945748
0.05	5	1000	0.8459458	0.5885955
0.05	5	5000	0.8370008	0.5692655
0.05	10	100	0.8409768	0.5543066
0.05	10	500	0.8494956	0.5951265
0.05	10	1000	0.8458041	0.5879926
0.05	10	5000	0.8365756	0.5683394
0.05	20	100	0.8409768	0.5543066
0.05	20	500	0.8510574	0.5995425
0.05	20	1000	0.8477924	0.5926063
0.05	20	5000	0.8360069	0.5662190
0.05	30	100	0.8411187	0.5547881
0.05	30	500	0.8510578	0.5997108
0.05	30	1000	0.8483610	0.5947162
0.05	30	5000	0.8350126	0.5646716
0.06	5	100	0.8408349	0.5564126
0.06	5	500	0.8475088	0.5910479
0.06	5	1000	0.8446681	0.5852050
0.06	5	5000	0.8370016	0.5697250
0.06	10	100	0.8408349	0.5564126
0.06	10	500	0.8490701	0.5950143
0.06	10	1000	0.8446681	0.5858142
0.06	10	5000	0.8316061	0.5555634
0.06	20	100	0.8411190	0.5572076
0.06	20	500	0.8499223	0.5974899
0.06	20	1000	0.8475079	0.5910982
0.06	20	5000	0.8365753	0.5682373
0.06	30	100	0.8406932	0.5556263
0.06	30	500	0.8482185	0.5926044
0.06	30	1000	0.8465143	0.5896928
0.06	30	5000	0.8357229	0.5667660

En este nuevo caso de la Tabla 33, se recomienda el uso de un *shrinkage* de 0.04, 500 árboles y 30 observaciones por nodo para alcanzar, en este caso, un accuracy de 0,8513419.
Se presentan los resultados en forma gráfica en la Figura 46.

Figura 46. Gráfico de resultados de 2° grilla de gradient boosting (R)

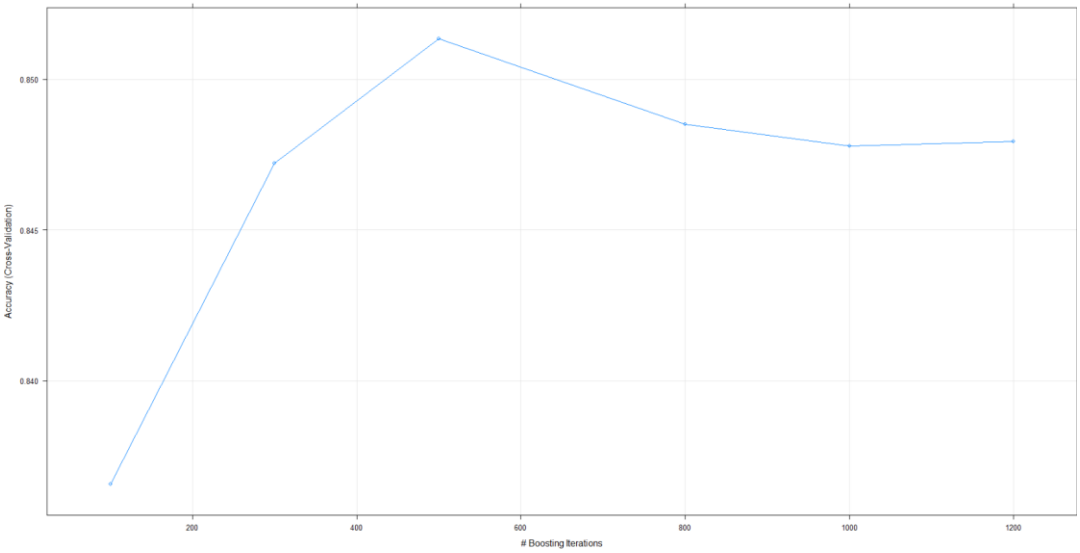


También se han realizado pruebas de sobreajuste por el número de árboles utilizados. En este caso, se establecen los hiperparámetros de *shrinkage* y número mínimo de observaciones en cada nodo en 0.04 y 30 respectivamente. Se varía el número de árboles entre los valores: 100, 300, 500, 800, 1000 y 1200. Los resultados se muestran en la Tabla 34.

Tabla 34. Resultados de estudio de número de árboles en gradient boosting (R)

N.TREES	ACCURACY	KAPPA
100	0.8365753	0.5384969
300	0.8472242	0.5824420
500	**0.8513419**	**0.5996028**
800	0.8485022	0.5942681
1000	0.8477923	0.5927556
1200	0.8479343	0.5936416

Figura 47. Gráfico de resultados de estudio de número de árboles en gradient boosting (R)



Con los valores de la Figura 47,se comprueba que el mejor número de árboles para la combinación precedente de hiperparámetros es de 500.

Al igual que en el caso de *random forest*, para *gradient boosting* se realiza un orden de importancia de las variables que se disponen. Se presentan en la tabla Tabla 35.

Tabla 35. Importancia de variables gradient boosting (R)

VARIABLE	REL. INF
CONTRACT_MONTH_TO_MONTH	37,18061892
TENURE_IN_MONTHS	12,90281847
NUMBER_OF_REFERRALS_1	10,09592236
AGE	7,42517462
NUMBER_OF_DEPENDENTS_0	7,34030062
MONTHLY_CHARGE	5,62686995
INTERNET_TYPE_FIBER_OPTIC	3,88281301
PAYMENT_METHOD_CREDIT_CARD	2,56752157
CONTRACT_TWO_YEAR	1,94025789
NUMBER_OF_REFERRALS_0	1,87691913
INTERNET_TYPE_NONE	1,26304085
AVG_MONTHLY_GB_DOWNLOAD	1,05138392
POPULATION	9,65574E-09
ONLINE_SECURITY	8,12001E-09
PAPERLESS_BILLING	7,78084E-09
TOTAL_CHARGES	7,68304E-09
TOTAL_LONG_DISTANCE_CHARGES	5,86642E-09
STREAMING_MUSIC	4,47192E-09
STREAMING_MOVIES	4,45269E-09
PREMIUM_TECH_SUPPORT	3,86741E-09
PAYMENT_METHOD_MAILED_CHECK	3,42747E-09
OFFER_OFFER_E	2,42662E-09
STREAMING_TV	2,34323E-09
AVG_MONTHLY_LONG_DISTANCE_CHARGE	1,57785E-09
NUMBER_OF_REFERRALS_8	1,24729E-09
CLTV	1,08869E-09
NUMBER_OF_REFERRALS_10	9,71264E-10
INTERNET_TYPE_CABLE	8,89743E-10
NUMBER_OF_REFERRALS_9	7,6415E-10
MULTIPLE_LINES	5,30811E-10
OFFER_OFFER_B	5,0945E-10
NUMBER_OF_REFERRALS_7	3,62641E-10
ONLINE_BACKUP	2,3481E-10
TOTAL_EXTRA_DATA_CHARGES	1,21973E-10
NUMBER_OF_REFERRALS_6	6,95244E-11
TOTAL_REFUNDS	0
OFFER_NONE	0
OFFER_OFFER_A	0
OFFER_OFFER_C	0
OFFER_OFFER_D	0

CONTRACT_ONE_YEAR	0
INTERNET_TYPE_DSL	0
PAYMENT_METHOD_BANK_WITHDRAWAL	0
DEVICE_PROTECTION_PLAN	0
GENDER_FEMALE	0
MARRIED	0
NUMBER_OF_DEPENDENTS_1	0
NUMBER_OF_DEPENDENTS_2	0
NUMBER_OF_DEPENDENTS_3	0
NUMBER_OF_REFERRALS_2	0
NUMBER_OF_REFERRALS_3	0
NUMBER_OF_REFERRALS_4	0
NUMBER_OF_REFERRALS_5	0
PARTNER	0
PHONE_SERVICE	0
UNLIMITED_DATA	0

Es visible en la Tabla 35 que solo 12 variables son realmente importantes para este algoritmo, siendo estas las primeras de la tabla. Específicamente son:

- Contract_Month_to_Month
- Tenure_in_Months
- Number_of_Referrals_1
- Age
- Number_of_Dependents_0
- Monthly_Charge
- Internet_Type_Fiber_Optic
- Payment_Method_Credit_Card
- Contract_Two_Year
- Number_of_Referrals_0
- Internet_Type_None
- Avg_Monthly_GB_Download.

No es casualidad que estas variables sean relevantes, debido a que la mayoría de estas han aparecido en las selecciones previas realizadas.

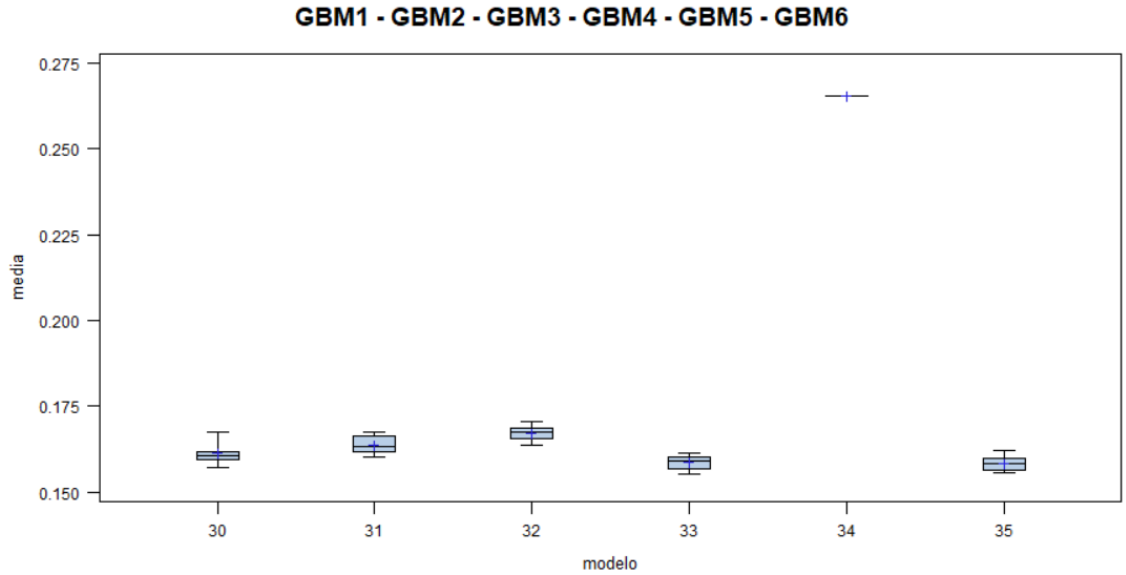
En SAS se utiliza la macro *%cruzadatreeboostbin* y se crean 6 modelos para realizar una primera aproximación con la configuración de la Tabla 36.

Tabla 36. Configuración hiperparámetros de gradient boosting (SAS)

MODELOS	**30**	**31**	**32**	**33**	**34**	**35**
LEAFSIZE	25	25	25	25	25	20
ITERACIONES	200	200	200	200	200	200
SHRINK	0.05	0.1	0.2	0.01	0.001	0.01
MAXBRANCH	4	4	4	4	4	4
MAXDEPTH	8	8	8	8	8	8
MINCATSIZE	15	15	15	15	5	10
MINOBS	20	20	20	20	20	20

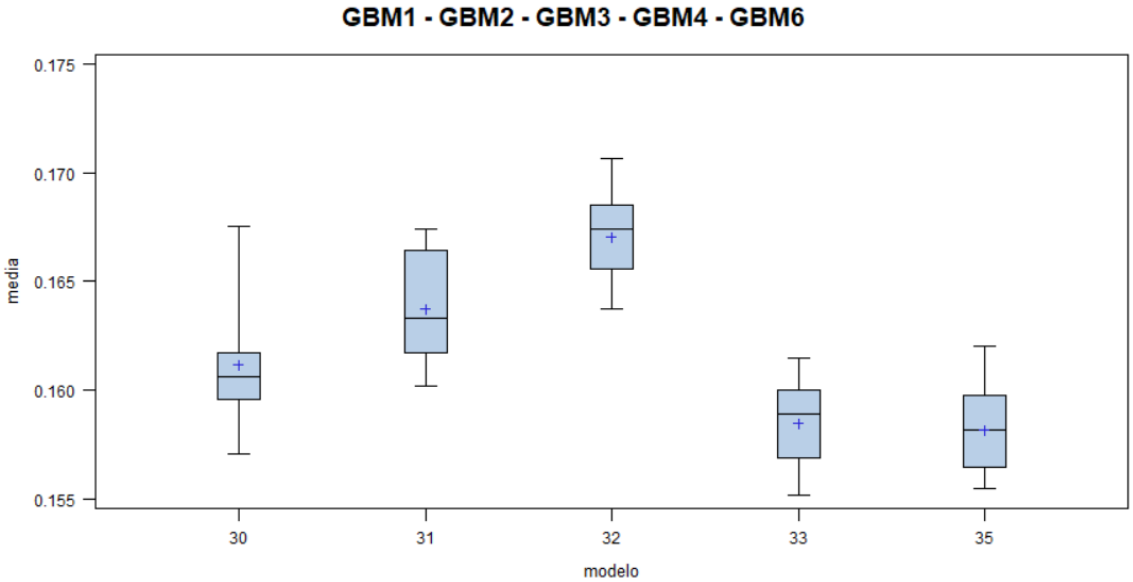
Los resultados se grafican con boxplots para cada uno de los modelos anteriores:

Figura 48. Tasa de fallos de gradient boosting (SAS)



Observando la Figura 48, el modelo 34 no es competitivo y es muy diferente al resto, a causa de que el parámetro *shrink* es demasiado pequeño.
Se procede a quitar este modelo del gráfico y se comprueba con mayor precisión la diferencia entre cada uno de los *gradient boostings* generados en la Figura 49.

Figura 49. Tasa de fallos de gradient boosting quitando "GBM5" (SAS)



Estos modelos son comparables con la regresión logística y las redes obtenidas en los apartados anteriormente.

Al igual que en R, en SAS se obtienen modelos con buen rendimiento con este algoritmo, particularmente los modelos 33 y 35.

12.4 Extreme gradient boosting (XGboost)

En R, se emplea el *grid* para definir diferentes valores para los hiperparámetros de varios modelos *extreme gradient boostings*:

- Min_child_weight: 5, 10 y 20
- Eta (shrinkage): 0.1, 0.05, 0.03, 0.01 y 0.001
- Nrounds (iteraciones): 100, 500, 1000 y 5000
- Max_depth: 6
- Gamma: 0

Una vez ejecutado el código con la configuración anterior y realizada la validación cruzada, se obtiene la Tabla 37 con los resultados:

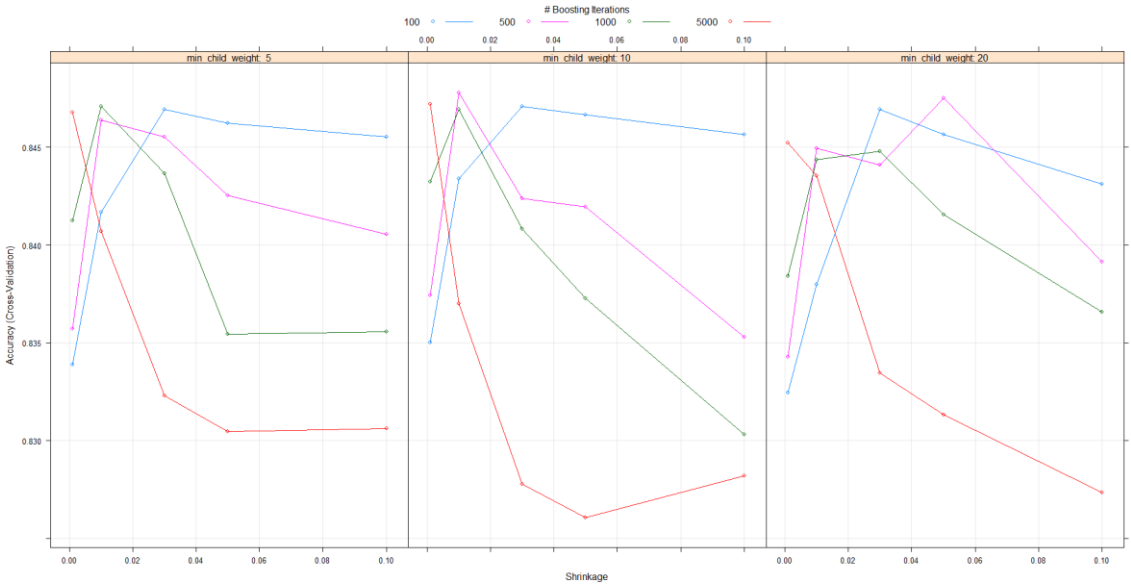
Tabla 37. Resultados de grilla de XGboost (R)

ETA	MIN_CHILD_WEIGHT	NROUNDS	ACCURACY	KAPPA
0.001	5	100	0.8338777	0.5464116
0.001	5	500	0.8357248	0.5491061
0.001	5	1000	0.8412620	0.5688162
0.001	5	5000	0.8467989	0.5891996
0.001	10	100	0.8350141	0.5507018
0.001	10	500	0.8374288	0.5552658
0.001	10	1000	0.8432493	0.5749830
0.001	10	5000	0.8472243	0.5912899
0.001	20	100	0.8324585	0.5365095
0.001	20	500	0.8343041	0.5447345
0.001	20	1000	0.8384215	0.5580822
0.001	20	5000	0.8452382	0.5870043
0.010	5	100	0.8416879	0.5695736
0.010	5	500	0.8463729	0.5882144
0.010	5	1000	0.8470824	0.5925002
0.010	5	5000	0.8406936	0.5781140
0.010	10	100	0.8433915	0.5762210
0.010	**10**	**500**	**0.8477931**	**0.5925918**
0.010	10	1000	0.8469406	0.5924200
0.010	10	5000	0.8370032	0.5662323
0.010	20	100	0.8379952	0.5570089
0.010	20	500	0.8449544	0.5858258
0.010	20	1000	0.8443852	0.5852604
0.010	20	5000	0.8435344	0.5846196
0.030	5	100	0.8469414	0.5865566
0.030	5	500	0.8455203	0.5891583
0.030	5	1000	0.8436750	0.5839283
0.030	5	5000	0.8323165	0.5546949
0.030	10	100	0.8470821	0.5886753
0.030	10	500	0.8423974	0.5807143
0.030	10	1000	0.8408359	0.5754437
0.030	10	5000	0.8277722	0.5431508
0.030	20	100	0.8469410	0.5875352

0.030	20	500	0.8441011	0.5849684
0.030	20	1000	0.8448120	0.5865806
0.030	20	5000	0.8334523	0.5575007
0.050	5	100	0.8462311	0.5878148
0.050	5	500	0.8425392	0.5810921
0.050	5	1000	0.8354411	0.5630753
0.050	5	5000	0.8304711	0.5502379
0.050	10	100	0.8466570	0.5896176
0.050	10	500	0.8419712	0.5792889
0.050	10	1000	0.8372862	0.5665609
0.050	10	5000	0.8260695	0.5382878
0.050	20	100	0.8456640	0.5878828
0.050	20	500	0.8475087	0.5944228
0.050	20	1000	0.8415455	0.5795922
0.050	20	5000	0.8313224	0.5543986
0.100	5	100	0.8455209	0.5902339
0.100	5	500	0.8405527	0.5769254
0.100	5	1000	0.8355826	0.5645964
0.100	5	5000	0.8306123	0.5520483
0.100	10	100	0.8456628	0.5890929
0.100	10	500	0.8352994	0.5622873
0.100	10	1000	0.8303306	0.5497062
0.100	10	5000	0.8281999	0.5455052
0.100	20	100	0.8431076	0.5818720
0.100	20	500	0.8391322	0.5717793
0.100	20	1000	0.8365766	0.5653358
0.100	20	5000	0.8273474	0.5437190

La Tabla 37 se puede plantear en formato gráfico como el que aprecia en la Figura 50.

Figura 50. Gráfico de resultados de grilla de XGboost (R)



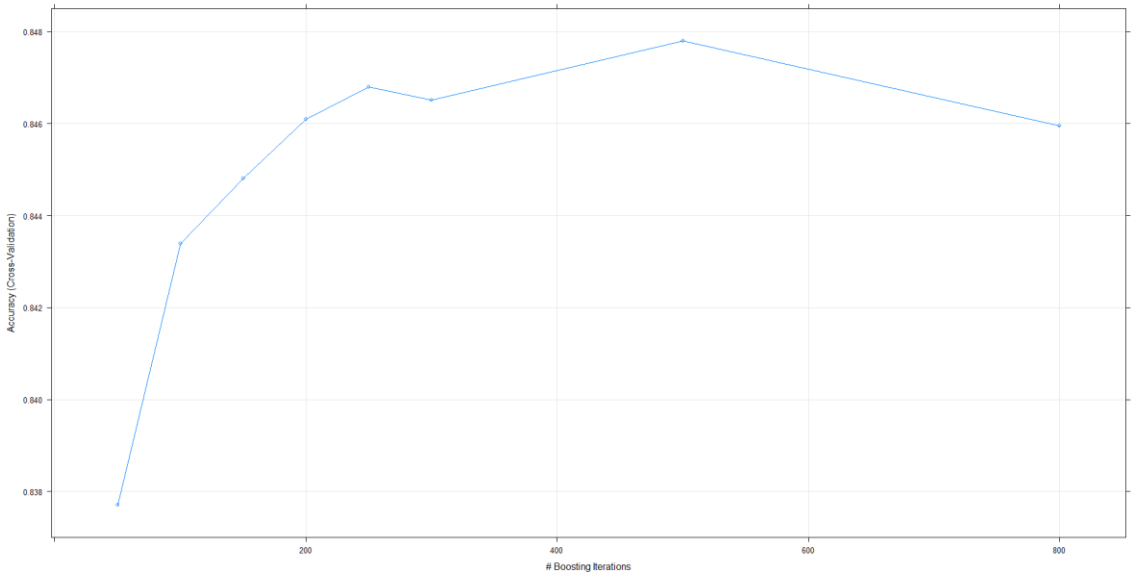
Con valores bajos de *shrinkage* se obtienen los valores más altos de accuracy.
Se utilizará el valor de 0,01 en adelante.

También se ha estudiado el número necesario de iteraciones para obtener un modelo cercano al óptimo con la siguiente configuración:

- Min_child_weight: 10
- Eta (shrinkage): 0.1
- Nrounds (iteraciones): 50, 100, 150, 200, 25, 300, 500, 800, 1200, 1500 y 2000
- Max_depth: 6
- Gamma: 0

Gráficamente, en la Figura 51, se pueden observar la variación de accuracy determinado por el aumento de las iteraciones sobre la configuración de xgboost anterior:

Figura 51. Gráfico de variación de accuracy por número de iteraciones de XGboost (R)



Se realiza un estudio del “early stopping” o búsqueda del valor óptimo de iteraciones, para los mismos hiperparámetros anteriores, pero utilizando dos semillas diferentes:

Figura 52. Gráfico de variación de accuracy por número de iteraciones de XGboost con cambio de semilla (R)

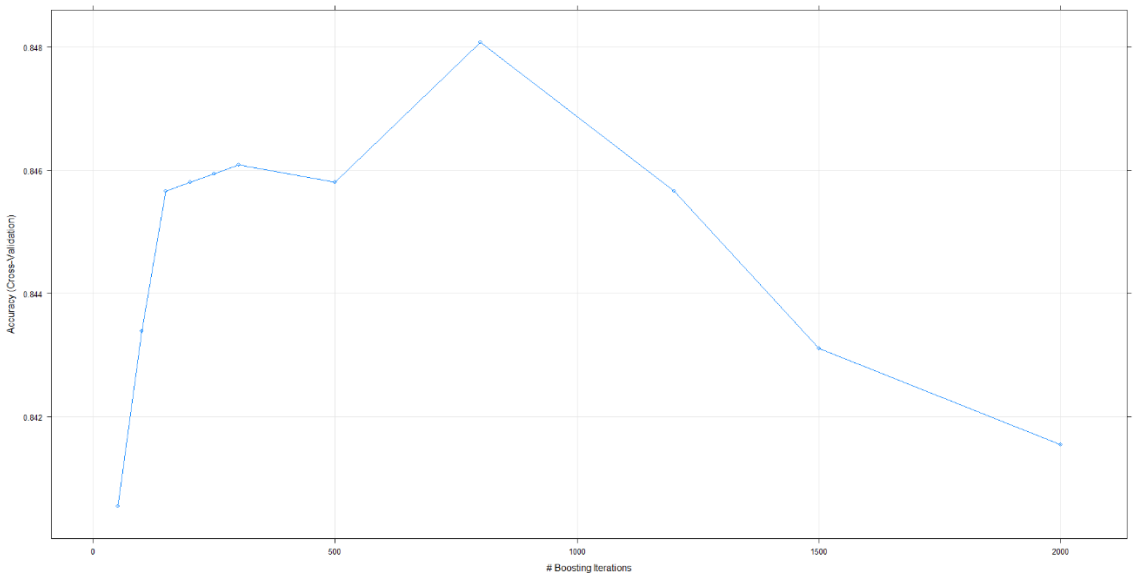
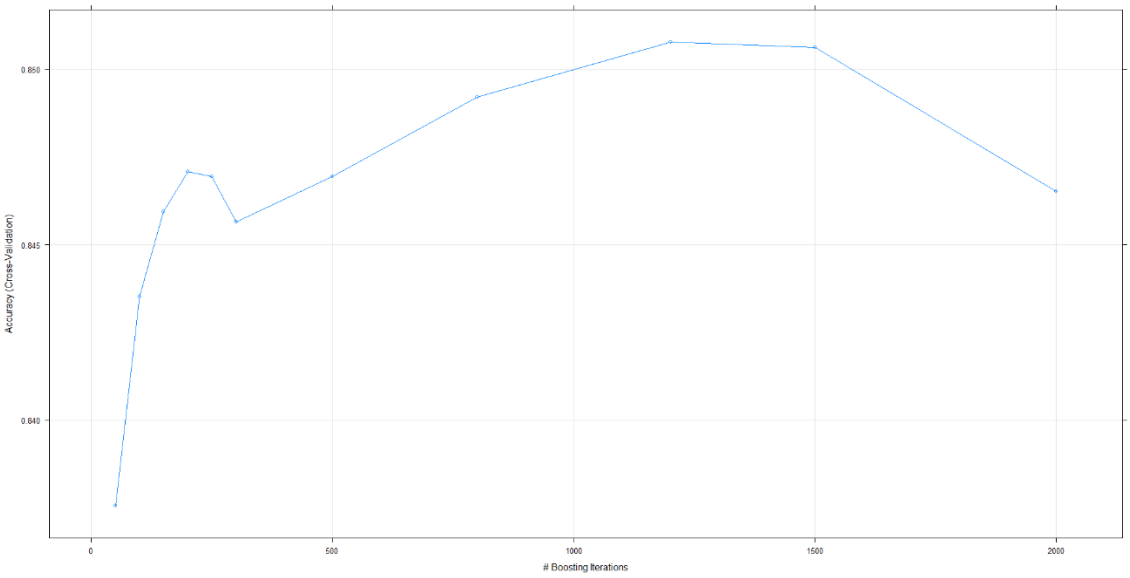


Figura 53. Gráfico de variación de accuracy por número de iteraciones de XGboost con segundo cambio de semilla (R)



Según los resultados de las Figuras Figura **52** y Figura **53**,parece que el mejor valor del hiperparámetro *nrounds* se encuentra entre 500 y 1200.

También se analiza las variables que se utilizan dentro del algoritmo para intentar predecir la probabilidad de abandono del cliente (“Churn”).
Solo se incluyen las de mayor importancia en la Tabla 38.

Tabla 38. Importancia de variables XGboost (R)

VARIABLE	OVERALL
CONTRACT_MONTH_TO_MONTH	100.000
AGE	23.599
TENURE_IN_MONTHS	20.236
NUMBER_OF_DEPENDENTS_0	19.410

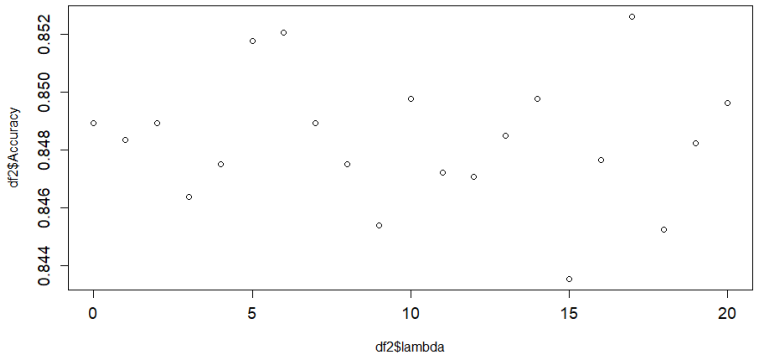
NUMBER_OF_REFERRALS_1	19.138
MONTHLY_CHARGE	15.842
NUMBER_OF_REFERRALS_0	6.984
PAYMENT_METHOD_CREDIT_CARD	6.732
TOTAL_LONG_DISTANCE_CHARGES	5.919
POPULATION	5.183
TOTAL_CHARGES	4.952
INTERNET_TYPE_FIBER_OPTIC	4.859
CLTV	3.894
CONTRACT_ONE_YEAR	3.661
AVG_MONTHLY_GB_DOWNLOAD	3.192
CONTRACT_TWO_YEAR	3.084
ONLINE_SECURITY	2.702
AVG_MONTHLY_LONG_DISTANCE_CHARGE	2.502
PAPERLESS_BILLING	2.112
STREAMING_MUSIC	2.064

12.4.1.1 Estudio de los parámetros de regularización

Utilizando un *xgboost* de 1000 iteraciones, un *shrinkage* de 0.01, una profundidad máxima de 6, un *min_child_weight* de 10, se realiza un proceso de iterativo sobre el parámetro de lambda, con valores de 0 a 20, aumentando de uno en uno.

Gráficamente se puede observar el desempeño del modelo:

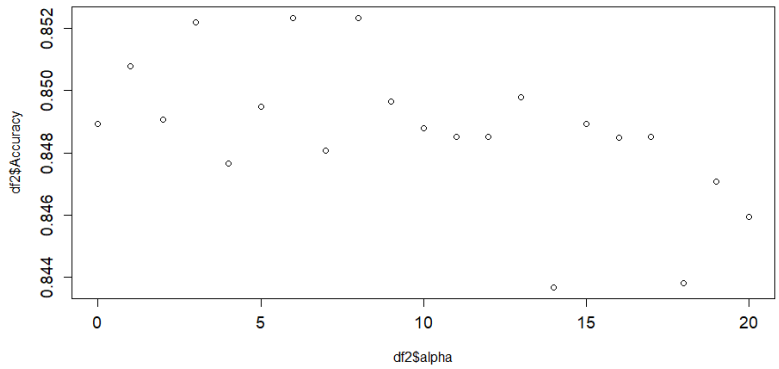
Figura 54. Resultados accuracy con variaciones en lambda de XGboost (R)



Aunque la ganancia sobre el accuracy no es demasiado grande, parece que valores lambda, en la Figura 54, de 5 o 6 podrían funcionar bastante bien para la optimización interna del algoritmo.

Se realiza un estudio similar para el parámetro alpha:

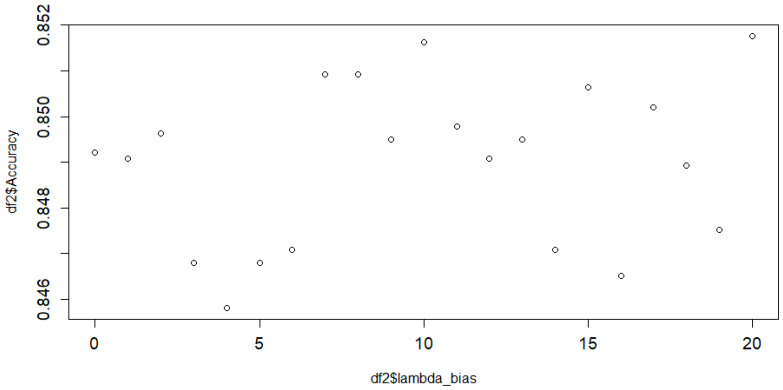
Figura 55. Resultados accuracy con variaciones en alpha de XGboost (R)



Observando los resultados en la Figura 55, los mejores valores parecen ser 3, 6 y 8.

Se analiza el parámetro de “lambda bias” también:

Figura 56. Resultados accuracy con variaciones en lambda bias de XGboost (R)



Con los valores de la Figura 56, parece que el mejor valor es el número 10 para este caso.

De las pruebas anteriores se han configurado diferentes modelos de xgboost con el objetivo de combinar los mejores valores de estos hiperparámetros y comparar los resultados de dichos modelos.

12.5 Support Vector Machine

En R, también se ha utilizado el *grid* de la librería *caret* para obtener los valores optimos de este algoritmo.

Se han utilizado diferentes valores del parámetro de penalización (C) para un modelo de SVM lineal:

- C: 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5 y 10

Los resultados de accuracy obtenidos se presentan en la Tabla 39.

Tabla 39. Resultados de grilla de SVM lineal (R)

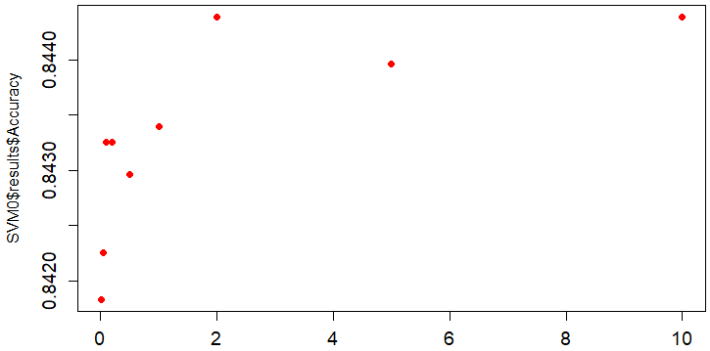
C	ACCURACY	KAPPA	ACCURACYSD	KAPPASD
0.01	0.8418307	0.5826898	0.005986878	0.01808504
0.05	0.8422570	0.5854947	0.006339270	0.01876469
0.10	0.8432507	0.5889438	0.005621542	0.01556776
0.20	0.8432505	0.5886846	0.005871212	0.01528053

0.50	0.8429663	0.5878946	0.004804144	0.01397541
1.00	0.8433926	0.5883386	0.005875798	0.01738200
2.00	**0.8443864**	**0.5901578**	**0.006069697**	**0.01807238**
5.00	0.8439595	0.5892683	0.004482483	0.01367913
10.00	0.8443857	0.5901645	0.004745303	0.01476666

El mayor valor de accuracy se obtiene con el valor 2 para el parámetro C.

Se realiza un gráfico (Figura 57) utilizando los valores de C y el accuracy obtenido:

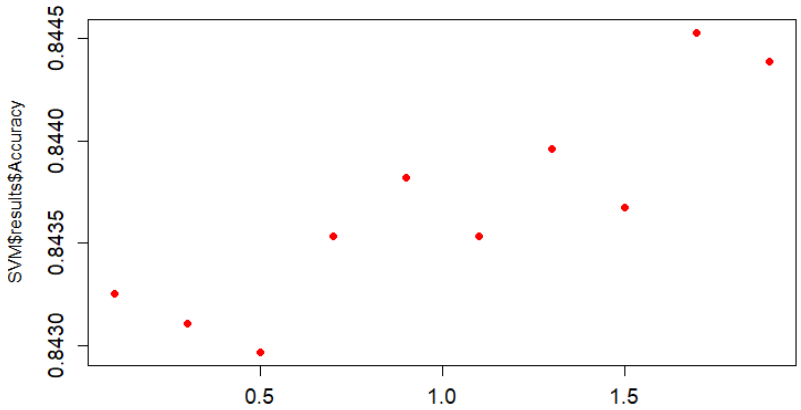
Figura 57. Gráfico de resultados de accuracy de SVM lineal (R)



Se observa en la Figura 57 que la variación del accuracy es muy leve entre todos los valores de C.

Aun así, se realiza una nueva prueba con valores de C entre 0.01 y 2.

Figura 58. Gráfico de resultados de accuracy de SVM lineal con valores reducidos del parámetro C (R)



En la Figura 58 la variación del accuracy se mantiene baja aun variando el parámetro C, si se observa los valores que se muestran de esta medida en el margen izquierdo del gráfico.

Por otro lado, se realiza una prueba con el algoritmo *SVM Polinomial* de la con los siguientes valores:

- C: 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5 y 10
- Degre (grado del polinomio): 2 y 3
- Scale: 0.1, 0.5, 1, 2 y 5

Con estas características se han obtenido los siguientes resultados:

Tabla 40. Resultados de grilla de SVM polinomial (R)

C	DEGREE	SCALE	ACCURACY	KAPPA	ACCURACY SD	KAPPA SD
0.01	**2**	**0.1**	**0.8436762**	**0.5794547**	**0.008138573**	**0.024929822**
0.05	2	0.1	0.8405516	0.5805888	0.005900494	0.017306683
0.10	2	0.1	0.8357231	0.5695921	0.004375160	0.011906644
0.20	2	0.1	0.8334514	0.5669935	0.002658782	0.006050225
0.50	2	0.1	0.8243620	0.5477143	0.008125802	0.016446514
1.00	2	0.1	0.8205291	0.5396317	0.006990671	0.013592525
2.00	2	0.1	0.8152764	0.5285767	0.003988563	0.007329929
5.00	2	0.1	0.8113027	0.5191812	0.002957495	0.010829565
10.00	2	0.1	0.8087456	0.5128503	0.003651696	0.011204350
0.01	2	0.5	0.8304697	0.5591656	0.003794484	0.009593170
0.05	2	0.5	0.8196769	0.5377329	0.006261355	0.013072229
0.10	2	0.5	0.8141403	0.5256976	0.004875818	0.008600405
0.20	2	0.5	0.8103089	0.5175389	0.004594045	0.014497764
0.50	2	0.5	0.8068987	0.5094001	0.006377644	0.017135417
1.00	2	0.5	0.8017878	0.4992635	0.007575386	0.021157804
2.00	2	0.5	0.7993731	0.4956347	0.006003316	0.016666384
5.00	2	0.5	0.7962485	0.4893778	0.008460784	0.020925970
10.00	2	0.5	0.7955395	0.4869435	0.006435507	0.018003142
0.01	2	1.0	0.8196768	0.5368516	0.009085112	0.018755806
0.05	2	1.0	0.8105929	0.5180841	0.004313663	0.014497617
0.10	2	1.0	0.8060482	0.5070219	0.003125529	0.012221590
0.20	2	1.0	0.8030654	0.5016953	0.007329994	0.021410853
0.50	2	1.0	0.7982369	0.4929713	0.007011430	0.017961365
1.00	2	1.0	0.7963918	0.4892670	0.005583273	0.014984593
2.00	2	1.0	0.7966744	0.4890677	0.009376670	0.020658412
5.00	2	1.0	0.7880143	0.4718979	0.005639119	0.014023035
10.00	2	1.0	0.7857409	0.4684871	0.010373745	0.020843709
0.01	2	2.0	0.8115859	0.5207271	0.003057444	0.009793557
0.05	2	2.0	0.8027818	0.5009956	0.006573737	0.019388165
0.10	2	2.0	0.7973854	0.4904889	0.006384911	0.016856519
0.20	2	2.0	0.7955393	0.4866196	0.006133590	0.015977951
0.50	2	2.0	0.7955390	0.4865628	0.008570073	0.018431831
1.00	2	2.0	0.7932680	0.4795740	0.005561005	0.012764094
2.00	2	2.0	0.7846053	0.4670141	0.009058617	0.019558570
5.00	2	2.0	0.7758013	0.4477646	0.013139903	0.025039734
10.00	2	2.0	0.7773610	0.4484374	0.016412958	0.035731068
0.01	2	5.0	0.7983806	0.4918278	0.004927166	0.015997136
0.05	2	5.0	0.7951131	0.4870487	0.007303174	0.017710081
0.10	2	5.0	0.7925577	0.4813822	0.006675252	0.014656300
0.20	2	5.0	0.7860274	0.4673222	0.003898186	0.008586931
0.50	2	5.0	0.7790689	0.4546951	0.007814331	0.021362537
1.00	2	5.0	0.7736724	0.4441432	0.011873480	0.024185490
2.00	2	5.0	0.7678505	0.4260570	0.014378271	0.035098795

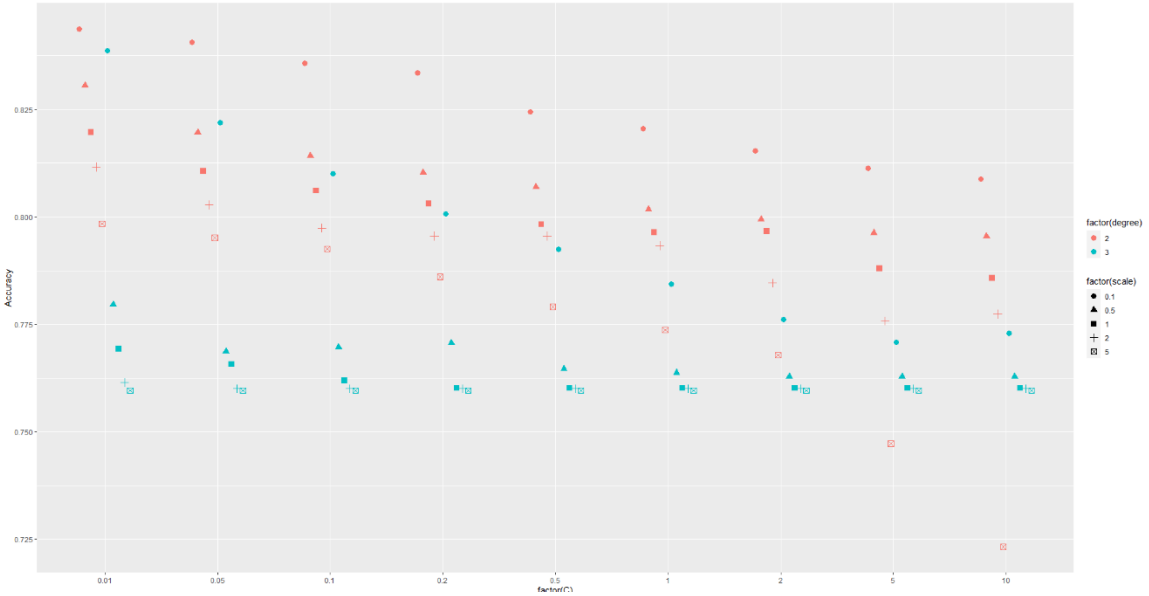
5.00	2	5.0	0.7472641	0.3658983	0.011184038	0.026735886
10.00	2	5.0	0.7232726	0.3036440	0.008584723	0.030570227
0.01	3	0.1	0.8387047	0.5783539	0.005862427	0.015670193
0.05	3	0.1	0.8219506	0.5431511	0.006477716	0.017620264
0.10	3	0.1	0.8100237	0.5144070	0.007796145	0.023850741
0.20	3	0.1	0.8006532	0.4929052	0.004964618	0.017526101
0.50	3	0.1	0.7924170	0.4751004	0.002619945	0.006718225
1.00	3	0.1	0.7843229	0.4562391	0.006233485	0.013492031
2.00	3	0.1	0.7760878	0.4369412	0.005769182	0.014726666
5.00	3	0.1	0.7708342	0.4268508	0.008011935	0.021408564
10.00	3	0.1	0.7729633	0.4338319	0.008908688	0.021429924
0.01	3	0.5	0.7796382	0.4449990	0.003475084	0.011567685
0.05	3	0.5	0.7687062	0.4223347	0.005429552	0.019060252
0.10	3	0.5	0.7696984	0.4251525	0.009574041	0.026103618
0.20	3	0.5	0.7706910	0.4297628	0.011450935	0.026777648
0.50	3	0.5	0.7645859	0.4147561	0.009519939	0.025064207
1.00	3	0.5	0.7637351	0.4138461	0.007398493	0.024705222
2.00	3	0.5	0.7628832	0.4121257	0.007953802	0.025641508
5.00	3	0.5	0.7628832	0.4121257	0.007953802	0.025641508
10.00	3	0.5	0.7628832	0.4121257	0.007953802	0.025641508
0.01	3	1.0	0.7692738	0.4258853	0.005105821	0.017261589
0.05	3	1.0	0.7657212	0.4181744	0.010170454	0.026152414
0.10	3	1.0	0.7618885	0.4103887	0.008184742	0.024320457
0.20	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
0.50	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
1.00	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
2.00	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
5.00	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
10.00	3	1.0	0.7601849	0.4057710	0.008136615	0.024834157
0.01	3	2.0	0.7614637	0.4088937	0.006585682	0.020521590
0.05	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
0.10	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
0.20	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
0.50	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
1.00	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
2.00	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
5.00	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
10.00	3	2.0	0.7600435	0.4058670	0.006882143	0.023498680
0.01	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
0.05	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
0.10	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
0.20	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
0.50	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
1.00	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
2.00	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
5.00	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047
10.00	3	5.0	0.7596186	0.4056900	0.005227114	0.020905047

Aun con todas estas combinaciones de la Tabla 40 de los parámetros C, grado de polinomio y scale no se obtiene una mejora sustancial aumentando la dimensión del mismo.

De hecho, como el mejor valor de accuracy obtenido corresponde al polinomio de grado 2, no parece que sea un problema a ser resuelto por un *support vector machine polinomial*.

Aun así, se grafica la evolución del acurracy para los polinomios de grado 2 y 3:

Figura 59. Gráfico de accuracy para SVM polinomiales de grado 2 y 3 (R)



Se observa en la Figura 59 que el accuracy disminuye a medida que se aumenta el parámetro de penalización (C) y el parámetro *scale*.

Con el objetivo de probar todas las variantes de este algoritmo, también se realiza una prueba con el *SVM Radial basis function (SVM RBF).*

Se han utilizado diferentes valores de C y de sigma:

- C: 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 y 30
- Sigma: 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 y 30

Tabla 41. Resultados de grilla de SVM RBF (R)

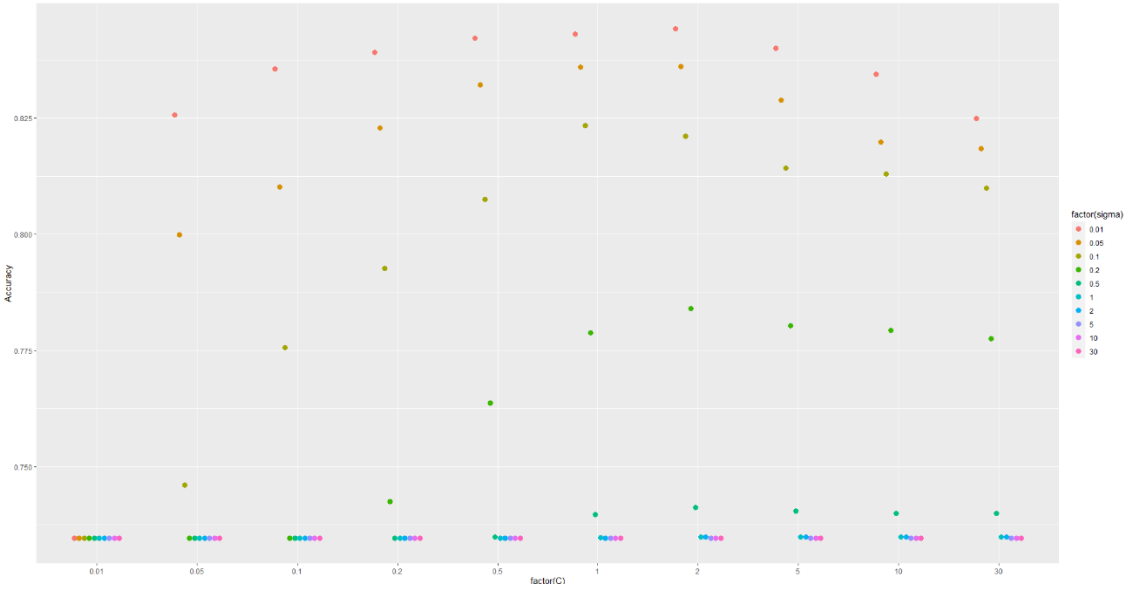
C	SIGMA	ACCURACY	KAPPA
0.01	0.01	0.7346302	0.000000000
0.01	0.05	0.7346302	0.000000000
0.01	0.10	0.7346302	0.000000000
0.01	0.20	0.7346302	0.000000000
0.01	0.50	0.7346302	0.000000000
0.01	1.00	0.7346302	0.000000000
0.01	2.00	0.7346302	0.000000000
0.01	5.00	0.7346302	0.000000000

0.01	10.00	0.7346302	0.000000000
0.01	30.00	0.7346302	0.000000000
0.05	0.01	0.8256437	0.502013971
0.05	0.05	0.7999443	0.369002347
0.05	0.10	0.7459893	0.072600923
0.05	0.20	0.7346302	0.000000000
0.05	0.50	0.7346302	0.000000000
0.05	1.00	0.7346302	0.000000000
0.05	2.00	0.7346302	0.000000000
0.05	5.00	0.7346302	0.000000000
0.05	10.00	0.7346302	0.000000000
0.05	30.00	0.7346302	0.000000000
0.10	0.01	0.8355830	0.544631578
0.10	0.05	0.8101686	0.421764469
0.10	0.10	0.7756651	0.241553969
0.10	0.20	0.7346300	0.001001812
0.10	0.50	0.7346302	0.000000000
0.10	1.00	0.7346302	0.000000000
0.10	2.00	0.7346302	0.000000000
0.10	5.00	0.7346302	0.000000000
0.10	10.00	0.7346302	0.000000000
0.10	30.00	0.7346302	0.000000000
0.20	0.01	0.8391323	0.562999047
0.20	0.05	0.8229462	0.482604347
0.20	0.10	0.7927032	0.328213472
0.20	0.20	0.7424399	0.051103741
0.20	0.50	0.7346302	0.000000000
0.20	1.00	0.7346302	0.000000000
0.20	2.00	0.7346302	0.000000000
0.20	5.00	0.7346302	0.000000000
0.20	10.00	0.7346302	0.000000000
0.20	30.00	0.7346302	0.000000000
0.50	0.01	0.8422556	0.576228809
0.50	0.05	0.8321753	0.529630498
0.50	0.10	0.8074702	0.411574606
0.50	0.20	0.7637393	0.176576182
0.50	0.50	0.7349141	0.001569606
0.50	1.00	0.7346302	0.000000000
0.50	2.00	0.7346302	0.000000000
0.50	5.00	0.7346302	0.000000000
0.50	10.00	0.7346302	0.000000000
0.50	30.00	0.7346302	0.000000000
1.00	0.01	0.8431072	0.581656331
1.00	0.05	0.8360093	0.551551250
1.00	0.10	0.8233742	0.483910264
1.00	0.20	0.7787898	0.263348717
1.00	0.50	0.7397420	0.033805233
1.00	1.00	0.7347724	0.002288892

1.00	2.00	0.7346302	0.000000000
1.00	5.00	0.7346302	0.000000000
1.00	10.00	0.7346302	0.000000000
1.00	30.00	0.7346302	0.000000000
2.00	**0.01**	**0.8442440**	**0.585862198**
2.00	0.05	0.8361525	0.560258895
2.00	0.10	0.8211019	0.487660502
2.00	0.20	0.7840437	0.301922788
2.00	0.50	0.7411631	0.058114027
2.00	1.00	0.7349147	0.011490218
2.00	2.00	0.7349142	0.004063897
2.00	5.00	0.7346302	0.000000000
2.00	10.00	0.7346302	0.000000000
2.00	30.00	0.7346302	0.000000000
5.00	0.01	0.8399852	0.579919540
5.00	0.05	0.8289117	0.541883837
5.00	0.10	0.8142866	0.466094786
5.00	0.20	0.7803528	0.290123933
5.00	0.50	0.7404533	0.058459931
5.00	1.00	0.7349147	0.012461150
5.00	2.00	0.7349142	0.004063897
5.00	5.00	0.7346302	0.000000000
5.00	10.00	0.7346302	0.000000000
5.00	30.00	0.7346302	0.000000000
10.00	0.01	0.8344487	0.568394246
10.00	0.05	0.8198235	0.516572584
10.00	0.10	0.8130090	0.462853553
10.00	0.20	0.7793589	0.288819449
10.00	0.50	0.7398853	0.057314641
10.00	1.00	0.7349147	0.012461150
10.00	2.00	0.7349142	0.004063897
10.00	5.00	0.7346302	0.000000000
10.00	10.00	0.7346302	0.000000000
10.00	30.00	0.7346302	0.000000000
30.00	0.01	0.8249354	0.548680316
30.00	0.05	0.8184043	0.512607969
30.00	0.10	0.8098839	0.454905721
30.00	0.20	0.7775135	0.284239278
30.00	0.50	0.7398853	0.057314641
30.00	1.00	0.7349147	0.012461150
30.00	2.00	0.7349142	0.004063897
30.00	5.00	0.7346302	0.000000000
30.00	10.00	0.7346302	0.000000000
30.00	30.00	0.7346302	0.000000000

Se observan los resultados de la Tabla 41**Tabla 41. Resultados de grilla de SVM RBF (R)** obtenidos en forma de gráfico:

Figura 60. Accurary de variaciones de SVM RBF (R)



Con los valores de Figura 60, el mejor caso de *SVM Radial* es el que utiliza un parámetro C igual a 2 y un sigma de 0.01.
Incluso con un valor de sigma por encima de 0.2 se obtienen valores similares de accuracy.

En SAS, para la creación de modelos utilizando *support vector machine* se utiliza la macro *%cruzadaSVMbin*:

Se realizan 4 modelos con las configuraciones de la Tabla 42:

Tabla 42. Configuración de hiperparámetros de SVM (SAS)

MODELOS	**40**	**41**	**42**	**43**
KERNEL	Rbf	Polynom	Polynom	Lineal
K_PAR	10	2	3	-
C	10	10	10	10

Los modelos 40, 41 y 42 no pudieron ser generados. Existe un error en SAS que no ha permitido construirlos y es el siguiente:

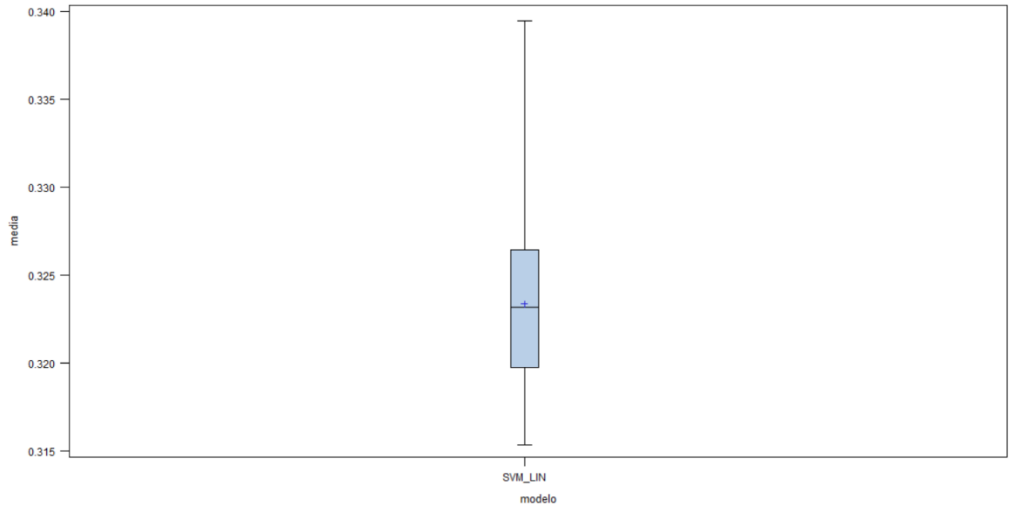
```
NOTE: Records processed = 5282   Memory used = 511K.
NOTE: There were 5282 observations read from the data set WORK.TRES.
NOTE: The data set WORK.CUA has 5282 observations and 32 variables.
NOTE: PROCEDURE DMDB used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds


5282 records read
NOTE: A parameter larger than one for the RBF function may result in floating point overflows.
NOTE: Method  LSSVM is very unusual and not very efficient for 5282 observations and 70 variables.
5282 records read
5282 records read
WARNING: For large problems like this, the iterative CG method is preferred.
WARNING: Problem too large for storing the kernel matrix, you must use the iterative CG method.
ERROR: Estimation cannot be performed.
NOTE: The SAS System stopped processing this step because of errors.
WARNING: The data set WORK.SAL6 may be incomplete.  When this step was stopped there were 0
         observations and 43 variables.
WARNING: Data set WORK.SAL6 was not replaced because this step was stopped.
```

Se produce el mismo error para el SVM polinomial.
Pese a que se han realizado intentos con menos variables, otros solo con las variables continuas o solo con las categóricas, no ha sido posible construirlos.

Es por ello, que a continuación solo se presenta el modelo obtenido con SVM lineal:

Figura 61. Tasa de fallos de SVM lineal (SAS)



En la Figura 61 se observa que el modelo 43 no es competitivo a la regresión logística, ha obtenido valores similares a los generados con *random forest*. Es por ello que no será incluido en la sección principal de este trabajo.

13 Anexo III – Valor óptimo de k

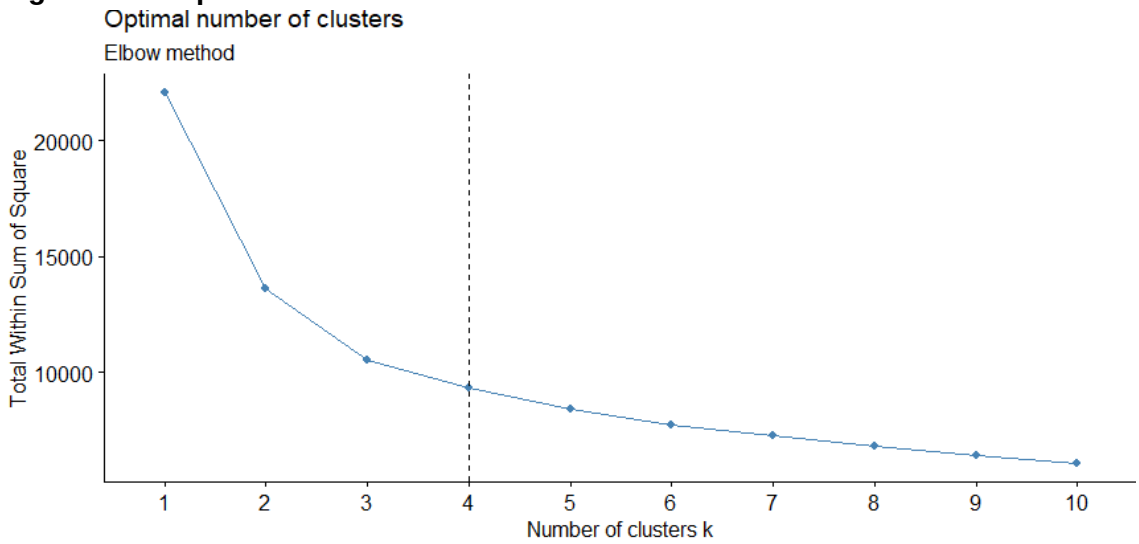
Con el propósito de complementar una decisión de negocio, en este caso de marketing, existen diferentes métodos y pruebas que permiten establecer el valor óptimo de k grupos a determinar para un conjunto de datos específico.

En R existen diferentes librerías que emplean estos métodos para encontrar el mejor valor de k, para luego utilizarlo como parámetro del algoritmo k-means.

Utilizando la librería *factoextra* se han utilizado 3 métodos para encontrar el mejor valor de k para los datos del presente trabajo, estos son los métodos: elbow, silhouette, y gap statistic.

El resultado de k óptimo de cada método es:

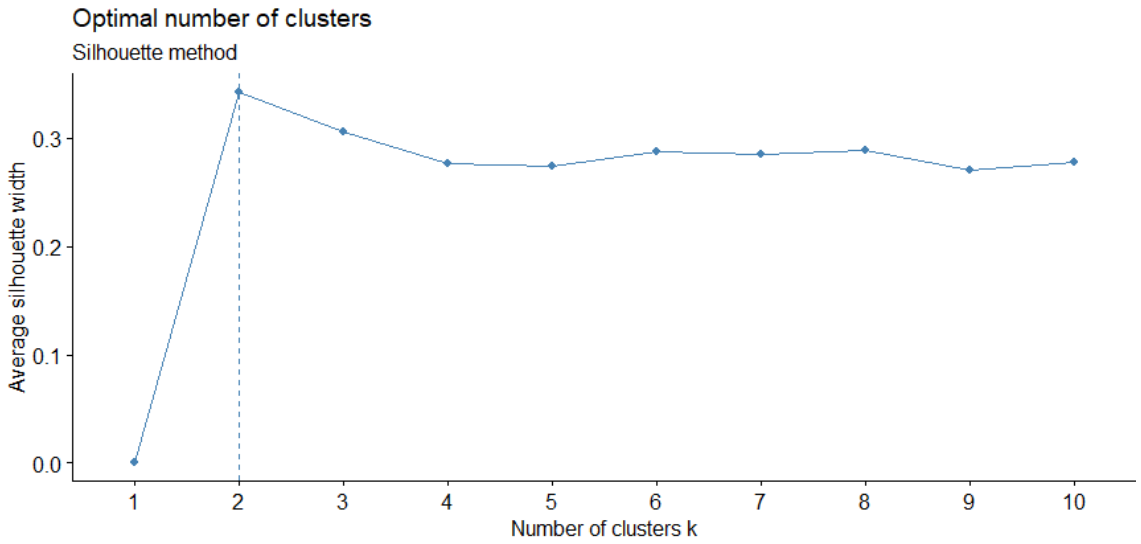
Figura 62. K óptimo método "elbow"



En el caso de la Figura 62, se encuentra que el valor óptimo de k es de 4.

El método “elbow” (método del codo) busca un valor óptimo de k grupos a partir de la optimización de la suma de cuadrados dentro de cada grupo. Es un método gráfico, en el cual se visualiza el valor de la suma de cuadrados dentro de cada grupo para diferentes valores de k y se selecciona el número de clusters que genera el “codo” de la curva, o un punto de inflexión. Dicho de otra forma, se selecciona el valor de k que si se le añadiese un valor más no generase una mejora sustancial.

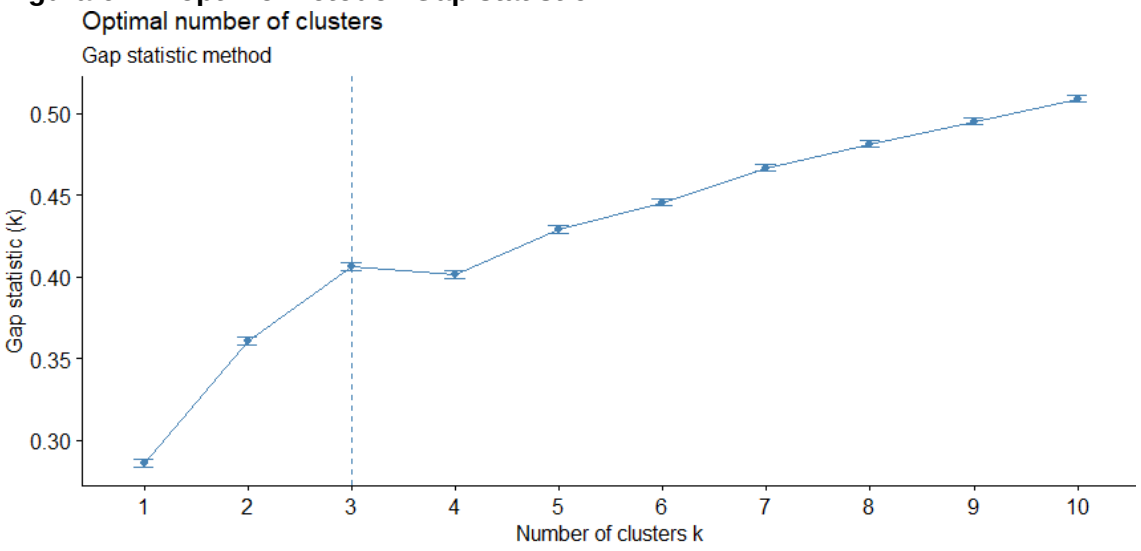
Figura 63. K óptimo método "silhouette"



El valor óptimo de k, por el máximo valor de “average silhouette width” obtenido (Figura 63), es de 2.

El “average silhoutte width” o “método de la silueta” mide la calidad de cada uno de los grupos generados. Básicamente, mide la distancia de separación de los clusters, es decir, indica que tan cerca se encuentran los puntos de un cluster con respecto a otro.

Figura 64. K óptimo método "Gap statistic"



Utilizando el método “gap statistic” se percibe que el mejor valor de cluster es de 3 (Figura 64).

El método “gap” (Tibshirani et al., 2001) o “brecha” compara la varianza total de cada cluster para diferentes valores de k, frente al valor esperado acorde a una distribución uniforme de referencia.

Con la libería *NbClust* se utiliza la distancia euclídea, con k-means y el método “silhouette” también se encuentra que el mejor valor de cluster es de 4 grupos.

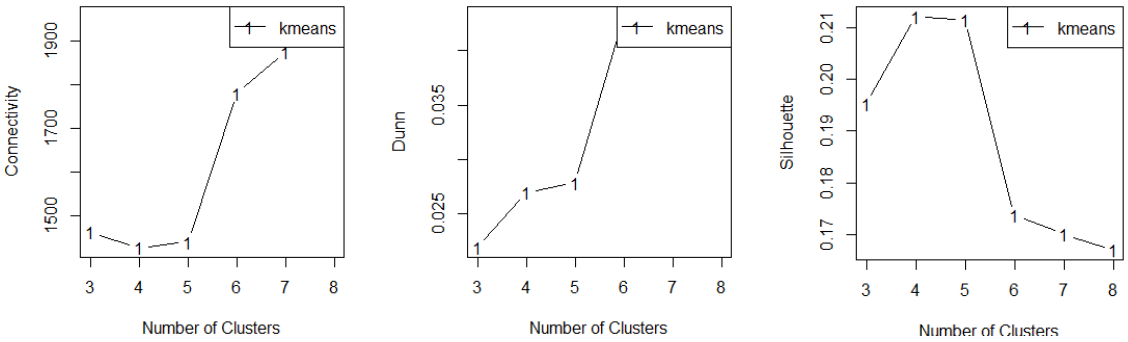
Por otro lado, utilizando la librería *clValid,* estableciendo pruebas de tres a ocho clusters, se puede observar el valor óptimo de con diferentes métodos (Tabla 43):

Tabla 43. K óptimos con clValid

OPTIMALSCORES	SCORE	METHOD	CLUSTERS
CONNECTIVITY	1,43E+03	kmeans	4
DUNN	4,32E-02	kmeans	6
SILHOUETTE	2,12E-01	kmeans	4

En forma gráfica:

Figura 65. K óptimos con clValid



Tiendo en cuenta la Figura 65, en dos de las tres pruebas se obtiene que el valor óptimo de k también es de cuatro.

14 Anexo IV – Código utilizado en SAS base

14.1 Macros

```
/* randomselectlog */
%macro
randomselectlog(data=,listclass=,vardepen=,modelo=,sinicio=,sfinal=,fr
acciontrain=,directorio=);
options nocenter linesize=256;
proc printto print="&directorio\kk.txt";run;
data;file "&directorio\cosa2.txt" ;run;
%do semilla=&sinicio %to &sfinal;
proc    surveyselect    data=&data    rate=&fracciontrain    out=sal1234
seed=&semilla;run;

%if &listclass ne %then %do;
ods output type3=parametros;
proc logistic data=sal1234;
   class &listclass;
    model &vardepen= &modelo/ selection=stepwise;
run;
data parametros;length effect $20. modelo $ 20000;retain modelo " ";set
parametros end=fin;effect=cat(' ',effect);
if   _n_   ne   1   then   modelo=catt(modelo,'   ',effect);if   fin   then
do;variable=modelo;output;end;
run;
%end;
%else %do;
ods output  Logistic.ParameterEstimates=parametros;
proc logistic data=sal1234;
    model &vardepen= &modelo/ selection=stepwise;
run;
%end;
ods graphics off;
ods html close;
data;file "&directorio\cosa2.txt" mod;set parametros;
%if &listclass ne %then %do; put variable @@;%end;
%else %do; if _n_ ne 1 then put variable @@;%end;
run;
%end;
proc printto ;run;
data todos;
infile "&directorio\cosa2.txt";
length efecto $ 400;
input efecto @@;
if efecto ne 'Intercept' then output;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;

data todos;
infile "&directorio\cosa2.txt";
length efecto $ 200;
input efecto $ &&;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;
data;set sal;put efecto;run;
%mend;

/* cruzadalogistica */
```

```
%macro
cruzadalogistica(archivo=,vardepen=,conti=,categor=,ngrupos=,sinicio=,
sfinal=,objetivo=tasafallos);
title ' ';
data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
  data dos;set &archivo;u=ranuni(&semilla);
  proc sort data=dos;by u;run;
  data dos (drop=nume);
  retain grupo 1;
  set dos nobs=nume;
  if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
  run;
  data fantasma;run;
  %do exclu=1 %to &ngrupos;
    data tres;set dos;if grupo ne &exclu then vardep=&vardepen*1;
    proc logistic data=tres noprint;/*<<<<<******SE PUEDE QUITAR EL
NOPRINT */
    %if (&categor ne) %then %do;class &categor;model vardep=&conti
&categor ;%end;
    %else %do;model vardep=&conti;%end;
    output out=sal p=predi;run;
    data sal2;set sal;pro=1-predi;if pro>0.5 then pre11=1; else
pre11=0;
    if grupo=&exclu then output;run;
    proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
    data estadisticos (drop=count percent pre11 &vardepen);
    retain vp vn fp fn suma 0;
    set sal3 nobs=nume;
    suma=suma+count;
    if pre11=0 and &vardepen=0 then vn=count;
    if pre11=0 and &vardepen=1 then fn=count;
    if pre11=1 and &vardepen=0 then fp=count;
    if pre11=1 and &vardepen=1 then vp=count;
    if _n_=nume then do;
    porcenVN=vn/suma;
    porcenFN=FN/suma;
    porcenVP=VP/suma;
    porcenFP=FP/suma;
    sensi=vp/(vp+fn);
    especif=vn/(vn+fp);
    tasafallos=1-(vp+vn)/suma;
    tasaciertos=1-tasafallos;
    precision=vp/(vp+fp);
    F_M=2*Sensi*Precision/(Sensi+Precision);
    output;
    end;
    run;

    data fantasma;set fantasma estadisticos;run;
  %end;
  proc means data=fantasma sum noprint;var &objetivo;
  output out=sumaresi sum=suma mean=media;
  run;
  data sumaresi;set sumaresi;semilla=&semilla;
  data final (keep=suma media semilla);set final sumaresi;if suma=. then
delete;run;
%end;
proc print data=final;run;
%mend;
```

```
/* cruzadabinarianeural */
%macro
cruzadabinarianeural(archivo=,vardepen=,conti=,categor=,ngrupos=,sinic
io=,sfinal=,nodos=,algo=,objetivo=,early=,acti=tanh,directorio=);
title ' ';
data final;run;
proc printto print="&directorio\basura.txt";

/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
  data dos;set &archivo;u=ranuni(&semilla);
  proc sort data=dos;by u;run;
  data dos (drop=nume);
  retain grupo 1;
  set dos nobs=nume;
  if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
  run;
  data fantasma;run;
  %do exclu=1 %to &ngrupos;

  data trestr tresval;
    set dos;if grupo ne &exclu then output trestr;else output tresval;
    PROC DMDB DATA=trestr dmdbcat=catatres;
    target &vardepen;
    var &conti;
    class &vardepen;
    %if &categor ne %then %do;class &categor &vardepen;%end;
    run;
    proc neural data=trestr dmdbcat=catatres random=789 ;
    input &conti;
    %if &categor ne %then %do;input &categor /level=nominal;%end;
    target &vardepen /level=nominal;
    hidden &nodos /act=&acti;/*<<<<<******PARA DATOS LINEALES ACT=LIN
(funciónde activación lineal)
    NORMALMENTE PARA DATOS NO LINEALES MEJOR ACT=TANH */
    /* A PARTIR DE AQUI SON ESPECIFICACIONES DE LA RED, SE PUEDEN
CAMBIAR O AÑADIR COMO PARÁMETROS */

    /*nloptions maxiter=500*/;
    netoptions randist=normal ranscale=0.15 random=15459;
    /* Si se desea hacer early stopping se pone prelim 0 y se marca
como comentario
    la línea prelim 15...*/
    /*prelim 0 */
    prelim 15 preiter=10 pretech=&algo;
    train maxiter=&early outest=mlpest technique=&algo;
    score data=tresval role=valid out=sal ;
    run;
    data sal2;set sal;pro=1-%str(p_&vardepen)0;if pro>0.5 then pre11=1;
else pre11=0;run;
         proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;

  data estadisticos (drop=count percent pre11 &vardepen);
    retain vp vn fp fn suma 0;
    set sal3 nobs=nume;
    suma=suma+count;
    if pre11=0 and &vardepen=0 then vn=count;
    if pre11=0 and &vardepen=1 then fn=count;
    if pre11=1 and &vardepen=0 then fp=count;
```

```
     if pre11=1 and &vardepen=1 then vp=count;
     if _n_=nume then do;
     porcenVN=vn/suma;
     porcenFN=FN/suma;
     porcenVP=VP/suma;
     porcenFP=FP/suma;
     sensi=vp/(vp+fn);
     especif=vn/(vn+fp);
     tasafallos=1-(vp+vn)/suma;
     tasaciertos=1-tasafallos;
     precision=vp/(vp+fp);
     F_M=2*Sensi*Precision/(Sensi+Precision);
     output;
     end;
     run;
   data fantasma;set fantasma estadisticos;run;
   %end;
   proc means data=fantasma sum noprint;var &objetivo;
   output out=sumaresi sum=suma mean=media;
   run;
   data sumaresi;set sumaresi;semilla=&semilla;
   data final (keep=suma media semilla);set final sumaresi;if suma=. then
delete;run;
%end;
proc printto ;
proc print data=final;run;
%mend;

/* neuralbinariabasica */
%macro
neuralbinariabasica(archivo=,listconti=,listclass=,vardep=,nodos=,cort
e=,semilla=,porcen=,algo=levmar);
title '';
data archivobase;set &archivo nobs=nume;ene=int(&porcen*nume);
call symput('ene',left(ene));
run;

proc sort data=archivobase;by &vardep;run;

proc   surveyselect   data=archivobase   out=muestra   outall   N=&ene
seed=&semilla;
/*si se quiere estratificacion en el muestreo quitar los comentarios en
strata*/
/* strata &vardep /alloc=proportional;*/run;
data  train  valida;set  muestra;if  selected=1  then  output  train;else
output valida;run;

PROC DMDB DATA=train dmdbcat=cataprueba;
target &vardep;
var &listconti;
class &listclass &vardep;
run;

%if &listclass ne %then %do;
proc neural data=train dmdbcat=cataprueba;
input  &listconti;
input &listclass /level=nominal;
target  &vardep /level=nominal;
hidden &nodos;
prelim 5;
train tech=&algo;
```

```
score data=valida out=salpredi outfit=salfit ;
run;
%end;

%else %do;
proc neural data=train dmdbcat=cataprueba;
input  &listconti;
target  &vardep /level=nominal;
hidden &nodos;
prelim 5;
train tech=&algo;
score data=valida out=salpredi outfit=salfit ;
run;
%end;

data salpredi;set salpredi;if p_&vardep.1>&corte/100 then predi1=1;else
predi1=0;run;
proc freq data=salpredi;tables predi1*&vardep/out=sal1;run;

/* Cálculo de estadísticos */

data estadisticos (drop=count percent predi1 &vardep);
retain vp vn fp fn suma 0;
set sal1 nobs=nume;
suma=suma+count;
if predi1=0 and &vardep=0 then vn=count;
if predi1=0 and &vardep=1 then fn=count;
if predi1=1 and &vardep=0 then fp=count;
if predi1=1 and &vardep=1 then vp=count;
if _n_=nume then do;
if vn=. then vn=0;if fn=. then fn=0;if vp=. then vp=0;if fp=. then fp=0;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
if vp=0 then precision=0;
if vp=0 then sensi=0;
if vn=0 then especif=0;
F_M=2*Sensi*Precision/(Sensi+Precision);
output;
end;
run;
proc print data=estadisticos;run;

%mend;

/* cruzadarandomforestbin */
%macro cruzadarandomforestbin(archivo=,vardep=,conti=,categor=,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=2,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=12340,sfinal=12345,objetivo=tasafallos);    /*OJO    A
CAMBIAR LA SEMILLA*/

data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
```

```
data dos;set &archivo;u=ranuni(&semilla);
  proc sort data=dos;by u;run;
  data dos ;
  retain grupo 1;
  set dos nobs=nume;
  if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
  run;


data fantasma;run;

%do exclu=1 %to &ngrupos;
data tres;set dos;if grupo ne &exclu then vardep=&vardep;


ods listing close;
proc hpforest data=tres
maxtrees=&maxtrees
vars_to_try=&variables
trainfraction=&porcenbag
leafsize=&tamhoja
maxdepth=&maxdepth
alpha=&pvalor
exhaustive=5000
missing=useinsearch ;
target vardep/level=nominal;
input &conti/level=interval;
%if (&categor ne) %then %do;
  input &categor/level=nominal;
  %end;
score out=salo;
run;
ods listing ;

data salo;merge salo tres;
if p_vardep1>0.5 then pre11=1;else pre11=0; /* hay que cambiar la
proporción */
if grupo=&exclu;
run;

  proc freq data=salo;tables pre11*&vardep/out=sal3;run;
    data estadisticos (drop=count percent pre11 &vardep);
    retain vp vn fp fn suma 0;
    set sal3 nobs=nume;
    suma=suma+count;
    if pre11=0 and &vardep=0 then vn=count;
    if pre11=0 and &vardep=1 then fn=count;
    if pre11=1 and &vardep=0 then fp=count;
    if pre11=1 and &vardep=1 then vp=count;
    if _n_=nume then do;
    porcenVN=vn/suma;
    porcenFN=FN/suma;
    porcenVP=VP/suma;
    porcenFP=FP/suma;
    sensi=vp/(vp+fn);
    especif=vn/(vn+fp);
    tasafallos=1-(vp+vn)/suma;
    tasaciertos=1-tasafallos;
    precision=vp/(vp+fp);
    F_M=2*Sensi*Precision/(Sensi+Precision);
```

```
     output;
     end;
     run;

data fantasma;set fantasma estadisticos;run;

%end;/* fin grupos */
  proc means data=fantasma sum noprint;var &objetivo;
  output out=sumaresi sum=suma mean=media;
  run;
  data sumaresi;set sumaresi;semilla=&semilla;
  data final (keep=suma media semilla);set final sumaresi;if suma=. then
delete;run;
%end;/* fin semillas validación cruzada repetida*/


proc print data=final;run;


%mend;


/* cruzadatreeboostingbin */
%macro
cruzadatreeboostbin(archivo=,vardepen=,conti=,categor=,ngrupos=,sinici
o=,sfinal=,leafsize=5,
iteraciones=100,shrink=0.01,maxbranch=2,maxdepth=4,mincatsize=15,minob
s=20,objetivo=tasafallos);
data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
  data dos;set &archivo;u=ranuni(&semilla);
  proc sort data=dos;by u;run;
  data dos ;
  retain grupo 1;
  set dos nobs=nume;
  if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
  run;
  data fantasma;run;
  %do exclu=1 %to &ngrupos;
    data tres;set dos;if grupo ne &exclu then vardep=&vardepen;

  proc treeboost data=tres
  exhaustive=1000 intervaldecimals=max
  leafsize=&leafsize iterations=&iteraciones maxbranch=&maxbranch
  maxdepth=&maxdepth     mincatsize=&mincatsize     missing=useinsearch
shrinkage=&shrink
  splitsize=&minobs;
  %if (&categor ne) %then %do;
  input &categor/level=nominal;
  %end;
  input &conti/level=interval;
  target vardep /level=binary;
  save fit=iteraciones importance=impor model=modelo rules=reglas;
  subseries largest;
  score out=sal;

  data  sal2;set  sal;pro=1-p_vardep0;if  pro>0.5  then  pre11=1;  else
pre11=0;/* hay que cambiar la proporción */
     if grupo=&exclu then output;run;
     proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
```

```
     data estadisticos (drop=count percent pre11 &vardepen);
     retain vp vn fp fn suma 0;
     set sal3 nobs=nume;
     suma=suma+count;
     if pre11=0 and &vardepen=0 then vn=count;
     if pre11=0 and &vardepen=1 then fn=count;
     if pre11=1 and &vardepen=0 then fp=count;
     if pre11=1 and &vardepen=1 then vp=count;
     if _n_=nume then do;
     porcenVN=vn/suma;
     porcenFN=FN/suma;
     porcenVP=VP/suma;
     porcenFP=FP/suma;
     sensi=vp/(vp+fn);
     especif=vn/(vn+fp);
     tasafallos=1-(vp+vn)/suma;
     tasaciertos=1-tasafallos;
     precision=vp/(vp+fp);
     F_M=2*Sensi*Precision/(Sensi+Precision);
     output;
     end;
     run;

     data fantasma;set fantasma estadisticos;run;
   %end;
   proc means data=fantasma sum noprint;var &objetivo;
   output out=sumaresi sum=suma mean=media;
   run;
   data sumaresi;set sumaresi;semilla=&semilla;
   data final (keep=suma media semilla);set final sumaresi;if suma=. then
delete;run;
%end;
proc print data=final;run;
%mend;


/* cruzadaSVMbin */
%macro
cruzadaSVMbin(archivo=,vardepen=,listclass=,listconti=,ngrupos=,sinici
o=,sfinal=,kernel=lineal,c=10,directorio=c:,
degree=2,k_par=1);
data final;run;
proc printto print="&directorio\ca.txt" log="&directorio\loga.txt";run;
%do semilla=&sinicio %to &sfinal;/*<<<<<******AQUI SE PUEDEN CAMBIAR LAS
SEMILLAS */
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos ;
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;

data fantasma;run;

%do exclu=1 %to &ngrupos;

data tres ;set dos;if grupo ne &exclu then vardep=&vardepen;run;

/*************************************************************/
/* SVM */
```

```
/*************************************************************/
%if &kernel=lineal %then %do;
proc hpsvm data=tres;
%if (&listclass ne) %then %do;
input &listclass/ level=nominal;
%end;
input &listconti / level=interval;
target vardep;
penalty C=&c;
output out=sal6;
run;

data sal6;merge tres sal6;run;

data salbis(keep=&vardepen p_vardep1 grupo vardep pre1);set sal6;
if grupo=&exclu;
predi1=p_vardep1;pre1=I_vardep;
run;
%end;

%else %if &kernel=polynom %then %do;

proc hpsvm data=tres;
input &listclass/ level=nominal;
input &listclass/ level=nominal;
input &listconti / level=interval;
target vardep;
kernel polynom /degree=&degree;
penalty C=&c;
output out=sal6;
run;

data sal6;merge tres sal6;run;

data salbis(keep=&vardepen p_vardep1 grupo vardep pre1);set sal6;
if grupo=&exclu;
predi1=p_vardep1;pre1=I_vardep;
run;

%end;

%else %if &kernel=RBF %then %do;

proc hpsvm data=tres method=activeset;
input &listclass/ level=nominal;
input &listconti / level=interval;
target vardep;
kernel RBF /k_par=&k_par;
penalty C=&c;
output out=sal6;
run;

data sal6;merge tres sal6;run;

data salbis(keep=&vardepen p_vardep1 grupo vardep pre1);set sal6;
if grupo=&exclu;
predi1=p_vardep1;pre1=I_vardep;
run;

%end;
```

```
data salbos;run;
proc freq data=salbis noprint;tables pre1*&vardepen /out=salconfu;run;
data confu1 (keep=tasa1);retain buenos 0 malos 0;set salconfu nobs=nume;
if &vardepen=pre1 then buenos=buenos+count;
if &vardepen ne pre1 then malos=malos+count;
if _n_=nume then do;tasa1=malos/(malos+buenos);output;end;
run;
data salbos;merge salbos confu1;run;
;

data fantasma;set fantasma salbos;run;

%end;
/* FIN GRUPOS */
proc means data=fantasma noprint;var tasa1;
output out=mediaresi mean=media;
run;
data mediaresi;set mediaresi;semilla=&semilla;run;
data final (keep=media semilla);set final mediaresi;if media=. then
delete;run;
%end;
proc printto; run;
proc print data=final;run;
%mend;


/* cruzadastackcon */
%macro cruzadastackcon
(archivo=,vardepen=,listclass=,listconti=,ngrupos=,seminicio=,semifina
l=,
nodos=10,algo=levmar,rediter=100,/*red*/
maxtrees=,vars_to_try=,trainfraction=,leafsize=,maxdepth=,/*random
forest */
bleafsize=,iterations=,bmaxbranch=,bmaxdepth=,shrinkage=,/*          g
boosting*/
kernel=lineal,c=10,degree=2,k_par=0.6 /*SVM*/);

data final;run;
*proc printto print='c:\ca.txt' log='c:\loga.txt';run;
%do  semilla=&seminicio  %to  &semifinal;/*<<<<<******AQUI  SE  PUEDEN
CAMBIAR LAS SEMILLAS */
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;

data dos (drop=nume);
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;

data fantasma;run;
data unionsalfin;run;
data unifin;run;

%do exclu=1 %to &ngrupos;

data   tres;set   dos;semilla=&semilla;if   grupo   ne   &exclu   then
vardep=&vardepen*1;run;


/*************************************************************/
```

```
/* LOGISTICA */
proc logistic data=tres noprint;/*<<<<<******SE PUEDE QUITAR EL NOPRINT
*/
class &listclass;
model vardep=&listconti &listclass;
score out=saco;
;run;
/*************************************************************/

data sal1 (drop=p_1);set saco;predi1=p_1;run;

/*************************************************************/
/*RED */
PROC DMDB DATA=tres dmdbcat=catatres;
target vardep ;
var &listconti;
class vardep &listclass;
;run;

proc neural data=tres dmdbcat=catatres ;
input &listconti;
input &listclass /level=nominal;
target vardep/ id=o level=nominal;
hidden &nodos/ id=h act=tanh;
netoptions randist=normal ranscale=0.15 random=15459;
prelim 0 preiter=10 ;
/*prelim 15 preiter=10 ;*/
train maxiter=&rediter technique=&algo;
score data=tres out=salred;
run;

data sal2 (keep=&vardepen predi2 grupo vardep semilla);set
salred;predi2=p_vardep1;run;

/*************************************************************/
/*RANDOM FOREST*/
/*************************************************************/

proc hpforest data=tres
maxtrees=&maxtrees                              vars_to_try=&vars_to_try
trainfraction=&trainfraction leafsize=&leafsize maxdepth=&maxdepth
exhaustive=5000
missing=useinsearch ;
target vardep/level=nominal;
input &listconti/level=interval;
input &listclass/level=nominal;
score out=salo;
run;

data sal3 (keep=&vardepen predi3 grupo vardep);set
salo;predi3=p_vardep1;run;

/*************************************************************/
/*GRADIENT BOOSTING */
/*************************************************************/

  proc treeboost data=tres
  exhaustive=1000 intervaldecimals=max
  leafsize=&bleafsize iterations=&iterations maxbranch=&bmaxbranch
  maxdepth=&bmaxdepth         mincatsize=10         missing=useinsearch
shrinkage=&shrinkage
```

```
   splitsize=10;
   input &listclass/level=nominal;
   input &listconti/level=interval;
   target vardep /level=binary;
   subseries largest;
   score out=salboost;
run;
data     sal4     (keep=&vardepen     predi4     grupo     vardep);set
salboost;predi4=p_vardep1;run;

/*************************************************************/
/* SVM */
/*************************************************************/
data tres ;set dos;if grupo ne &exclu then vardep=&vardepen;run;

/*************************************************************/
/* SVM */
/*************************************************************/
%if &kernel=lineal %then %do;
proc hpsvm data=tres;
input &listclass/ level=nominal;
input &listconti / level=interval;
target vardep;
penalty C=&c;
output out=sal5;
run;

data sal5;merge tres sal5;run;

data sal5(keep=predi5);set sal5;
predi5=p_vardep1;
run;

%end;

%else %if &kernel=polynom %then %do;

proc hpsvm data=tres;
input &listclass/ level=nominal;
input &listconti / level=interval;
target vardep;
kernel polynom /degree=&degree;
penalty C=&c;
output out=sal5;
run;

data sal5;merge tres sal5;run;

data sal5(keep=predi5);set sal5;
predi5=p_vardep1;
run;

%end;

%else %if &kernel=RBF %then %do;

proc hpsvm data=tres method=activeset;
input &listclass/ level=nominal;
input &listconti / level=interval;
target vardep;
kernel RBF /k_par=&k_par;
```

```
penalty C=&c;
output out=sal5;
run;

data sal5;merge tres sal5;run;

data sal5(keep=predi5);set sal5;
predi5=p_vardep1;
run;

%end;


/* PRUEBAS CON STACKING */
data unionsal (drop=ygorro);merge sal1 sal2 sal3 sal4 sal5;
predi6=(predi1+predi2)/2; /* RED -LOG */
predi7=(predi1+predi3)/2;/* RED -RFOR */
predi8=(predi1+predi4)/2;/* RED -BOOST*/
predi9=(predi2+predi3)/2;/* LOG-RFOR */
predi10=(predi2+predi4)/2;/* LOG-BOOST */
predi11=(predi3+predi4)/2;/* RFOR-BOOST */
predi12=(predi1+predi2+predi3)/3;/* RED -LOG-RFOR */
predi13=(predi1+predi2+predi4)/3;/* RED -LOG-BOOST*/
predi14=(predi1+predi3+predi4)/3;/* RED -RFOR-BOOST*/
predi15=(predi2+predi3+predi4)/3;/* LOG-RFOR-BOOST*/
predi16=(predi1+predi2+predi3+predi4)/4;/* RED-LOG-RFOR-BOOST*/
predi17=(predi1*0.2+predi2*0.1+predi3*0.5+predi4*0.2);/*  RED-LOG-RFOR-
BOOST ponderado*/
predi18=(predi1+predi5)/2; /* RED -SVM */
predi19=(predi3+predi5)/2; /* RFOR -SVM */
predi20=(predi2+predi5)/2; /* LOG-SVM */
predi21=(predi4+predi5)/2; /* BOOST-SVM */
predi22=(predi5+predi2+predi3)/3;/* SVM-LOG-RFOR */
predi23=(predi1+predi2+predi3+predi4+predi5)/4;/*   RED-LOG-RFOR-BOOST-
SVM*/
run;


data salfin (keep=&vardepen vardep predi1-predi23 grupo);set unionsal;if
grupo=&exclu then output;run;

data unionsalfin;set unionsalfin salfin;run;

data salbis (drop=i);
array predi{23};
array pre{23};
set salfin;
do i=1 to 23;
if predi{i}>0.5 then pre{i}=1;/* se puede cambiar la proporción */
if predi{i}<=0.5 then pre{i}=0;
end;
run;
data salbos;run;
%do j=1 %to 23;
proc freq data=salbis noprint;tables pre&j*&vardepen /out=salconfu;run;
data  confu&j  (keep=tasa&j);retain  buenos  0  malos  0;set  salconfu
nobs=nume;
if &vardepen=pre&j then buenos=buenos+count;
if &vardepen ne pre&j then malos=malos+count;
if _n_=nume then do;tasa&j=malos/(malos+buenos);output;end;
run;
```

```
data salbos;merge salbos confu&j;run;
;
%end;

data fantasma;set fantasma salbos;run;
%end;

/* FIN GRUPOS */


proc means data=fantasma noprint;var tasa1-tasa23;
output out=mediaresi mean=ase1-ase23 ;
run;
data mediaresi;set mediaresi;semilla=&semilla;run;
data final (keep=ase1-ase23 semilla);set final mediaresi;if ASE1=. then
delete;run;

data unifin;set unifin unionsalfin;run;

%end;
proc printto; run;
proc print data=final;run;
%mend;
```

14.2 Desarrollo

```
/* Se ejecutan todas las macros juntas */
%include'D:\Fede\Google Drive\Master Minería de Datos e Inteligencia de
Negocios\Materias\Tecnicas de Machine
Learning\Trabajos\Clasificación\SAS\Ensamblado\pack cruzadas binarias
incluido ensamblado 7.0.sas';

ods graphics off;

/* CARGA DE DATOS (XLSX) - Categoricas en numeros*/
proc import datafile = 'D:\Fede\Google Drive\Master Minería de Datos
e Inteligencia de Negocios\Materias\Tecnicas de Machine
Learning\Trabajos\Clasificación\churn_clean_CatNum.xlsx'
 out = churn
 dbms = xlsx;
run;
/* SELECCIÓN DE VARIABLES */
/* MODELOS SIN INTERACCIONES */
ods output type3=parametros;
proc logistic data=churn namelen=40 descending ;
class Contract Device_Protection_Plan Gender_Female Internet_Type
Married
Multiple_Lines Number_of_Dependents Number_of_Referrals Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data ;
model Churn = Age Avg_Monthly_GB_Download
Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge Population Tenure_in_Months Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds
Contract Device_Protection_Plan Gender_Female Internet_Type Married
Multiple_Lines Number_of_Dependents Number_of_Referrals Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
```

```
Unlimited_Data
/selection=stepwise;
run;quit;
data mode;length effect $20. modelo $ 20000;retain modelo " ";set
parametros end=fin;effect=cat(' ',effect);
if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then output;
run;
data ;set mode;put modelo;run;

/* CON LA OPCIÓN selection=score */
/* busca modelos tentativos y mejores con 1 a 18 variables*/
ods output bestsubsets=modelos;
proc logistic data=churn descending;
model                Churn=Age                Avg_Monthly_GB_Download
Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds
Gender_Female Multiple_Lines Online_Backup
Online_Security Paperless_Billing Partner Phone_Service
Premium_Tech_Support  Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data
/selection=score best=1 start=3 stop=18;
run;
data ;set modelos;put variablesinmodel;run;

/* Macro %randomselectlog */
%randomselectlog(data=churn,
listclass= Contract Device_Protection_Plan Gender_Female Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
vardepen=Churn,
modelo=Age   Avg_Monthly_GB_Download   Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds
Contract Device_Protection_Plan Gender_Female Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
sinicio=12345,sfinal=12380,fracciontrain=0.8,
directorio=D:\Fede\Google Drive\Master Minería de Datos e Inteligencia
de           Negocios\Materias\Tecnicas           de           Machine
Learning\Trabajos\Clasificación);


/* PROBAMOS LOS mejores MODELOS CON LOGÍSTICA */
%cruzadalogistica
(archivo=churn,vardepen=Churn,
conti=   Age   Avg_Monthly_GB_Download   Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
```

```
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365);
data final1;set final;modelo=1;


%cruzadalogistica
(archivo=churn,vardepen=Churn,
conti= Monthly_Charge Population Tenure_in_Months Total_Charges,
categor=Contract Married Number_of_Dependents Number_of_Referrals
Offer  Online_Backup  Online_Security  Paperless_Billing  Payment_Method
Phone_Service,
ngrupos=10,sinicio=12345,sfinal=12365);
data final3;set final;modelo=3;

%cruzadalogistica
(archivo=churn,vardepen=Churn,
conti= Population Tenure_in_Months Total_Charges Monthly_Charge,
categor=   Contract   Device_Protection_Plan   Internet_Type   Married
Number_of_Dependents    Number_of_Referrals    Offer    Online_Backup
Online_Security
Paperless_Billing  Payment_Method  Phone_Service  Premium_Tech_Support
Streaming_Movies Streaming_Music Streaming_TV,
ngrupos=10,sinicio=12345,sfinal=12365);
data final5;set final;modelo=5;

data union;set final1 final3 final5
proc boxplot data=union;plot media*modelo;run;
/* El modelo 1 es muy bueno*/

/* REDES NEURONALES */
/* LEVMAR */
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=churn,
   listconti=Age                                  Avg_Monthly_GB_Download
Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
   vardep=Churn,
   nodos=&nodos,corte=50,semilla=&semilla,porcen=0.80,algo=levmar);
   data                                              estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
```

```
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12355,inicionodos=3,finalnodos=11,incr
enodos=2);

/* Parece que cuantos menos nodos se utilizan es mejor. Utilizar 3 o 5
nodos*/


/* IGUAL CON BPROP */

%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=churn,
   listconti=Age                                  Avg_Monthly_GB_Download
Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,porcen=0.80,algo=b
prop mom=0.2 learn=0.1);
   data                                                estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12355,inicionodos=2,finalnodos=9,incre
nodos=1);
/* Parece que entre 3 y 9 nodos funciona bien */

%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=churn,
   listconti=Age                                  Avg_Monthly_GB_Download
Avg_Monthly_Long_Distance_Charge CLTV
```

```
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,porcen=0.80,algo=b
prop mom=0.3 learn=0.2);
   data                                              estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12355,inicionodos=2,finalnodos=13,incr
enodos=1);
/* Parece que entre 4 y 7 nodos funciona bien */


/* EARLY STOPPING */
%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442711,ocultos=3,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
```

```
vardep=Churn,porcen=0.80,semilla=442712,ocultos=3,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442713,ocultos=3,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442711,ocultos=5,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442712,ocultos=5,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
```

```
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442713,ocultos=5,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=levmar,acti=TAN
H);

%redneuronalbinaria(archivo=churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
   listclass=Contract       Device_Protection_Plan       Gender_Female
Internet_Type Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,

vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=levmar,acti=TAN
H);
```

```
/* COMPARAMOS LAS REDES CON LOGISTICA*/

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=5,algo=levmar,early=10,act
i=tanh);
data final6;set final;modelo=6;

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=7,algo=levmar,early=8,acti
=tanh);
data final7;set final;modelo=7;

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=7,algo=bprop        mom=0.3
learn=0.2);
data final8;set final;modelo=8;

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
```

```
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=5,algo=bprop       mom=0.3
learn=0.2);
data final9;set final;modelo=9;

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=7,algo=bprop       mom=0.2
learn=0.1);
data final10;set final;modelo=10;

%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,nodos=5,algo=bprop       mom=0.2
learn=0.1);
data final11;set final;modelo=11;

/* Todos los modelos */
ods graphics off;

title j=c '1-Log: All; 3-Log; 5-Log; 6-LV N5 E10 Tanh; 7-LV N7 E8 Tanh;
8-BP N7 M0,3 L0,2; 9-BP N5 M0,3 L0,2; 10-BP N7 M0,2 L0,1; 11-BP N5 M0,2
L0,1' ;
data union;set final1 final3 final5 final6 final7 final8 final9 final10
final11;
proc boxplot data=union;plot media*modelo;run;


/* PRUEBA DE RED NEURONAL CON FUNCION DE ACTIVACION LINEAL */
%cruzadabinarianeural(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
```

```
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=4,sinicio=12345,sfinal=12365,nodos=5,algo=levmar,early=10,acti
=lin);
data final12;set final;modelo=12;

title j=c '1-Log: All; 3-Log; 5-Log; 6-LV N5 E10 Tanh; 7-LV N7 E8 Tanh;
8-BP N7 M0,3 L0,2; 9-BP N5 M0,3 L0,2; 10-BP N7 M0,2 L0,1; 12-LV N5 E10
LIN' ;
data union;set final1 final3 final5 final6 final7 final8 final9 final10
final12;
proc boxplot data=union;plot media*modelo;run;
/* El modelo Levmar con Early Stopping y función de activación Lineal
es muy similar a la mejor regresión Logística */

title j=c '1-Log: All; 5-Log; 12-LV N5 E10 LIN' ;
data union;set final1 final5 final12;
proc boxplot data=union;plot media*modelo;run;


/* RANDOM FOREST */

/* Modelos variando MaxDepth y TamHoja*/

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=200,variables=30,porcenbag=0.80,maxbranch=4,tamhoja=30,maxdep
th=20,pvalor=0.1,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final21;set final;modelo=21;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=  Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
```

```
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=200,variables=25,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdep
th=15,pvalor=0.1,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final22;set final;modelo=22;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=  Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=200,variables=15,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdep
th=10,pvalor=0.1,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final23;set final;modelo=23;

ods graphics off;
data union;set final21 final22 final23 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=  Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=300,variables=15,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdep
th=8,pvalor=0.05,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final24;set final;modelo=24;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=  Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
```

```
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=400,variables=15,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdep
th=8,pvalor=0.01,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final25;set final;modelo=25;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=  Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=400,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdep
th=8,pvalor=0.01,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final26;set final;modelo=26;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=200,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=30,maxdep
th=10,pvalor=0.01,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final27;set final;modelo=27;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
```

```
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=150,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=8,pvalor=0.01,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final28;set final;modelo=28;

%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=200,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdept
h=6,pvalor=0.01,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final29;set final;modelo=29;


%cruzadarandomforestbin(
archivo=churn,
vardep=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
maxtrees=250,variables=15,porcenbag=0.80,maxbranch=2,tamhoja=20,maxdep
th=15,pvalor=0.05,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final281;set final;modelo=281;

proc boxplot data=final281;plot media*modelo;run;


ods graphics off;
data union;set final21 final22 final23 final24 final25 final26 final27
final28 final29;
proc boxplot data=union;plot media*modelo;run;

/* Mejor Random Forest: modelo28 */
```

```
/*
maxtrees=150,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=8,pvalor=0.01 */

/*---------------------------------------------------------------------
----------------------*/
/* GRADIENT BOOSTING */

/* MODELOS VARIANDO CTE Y FIJAMOS LEAFSIZE Y DEPTH COMO EN RANDOM
FOREST*/

%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
leafsize=25,
iteraciones=200,shrink=0.05,maxbranch=4,maxdepth=8,mincatsize=15,minob
s=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final30;set final;modelo=30;

%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
leafsize=25,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=8,mincatsize=15,minobs
=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final31;set final;modelo=31;

%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
```

```
Unlimited_Data,
leafsize=25,
iteraciones=200,shrink=0.2,maxbranch=4,maxdepth=8,mincatsize=15,minobs
=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final32;set final;modelo=32;

title j=c 'RF - GBM1 - GBM2 - GBM3';
data union;set final28 final30 final31 final32 ;
proc boxplot data=union;plot media*modelo;run;
/* Los 3 GBM superan ampliamente a RF*/

title j=c 'GBM1 - GBM2 - GBM3';
data union;set final30 final31 final32 ;
proc boxplot data=union;plot media*modelo;run;

/* se realizan variaciones sobre el mejor GBM*/
%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
leafsize=25,
iteraciones=200,shrink=0.01,maxbranch=4,maxdepth=8,mincatsize=15,minob
s=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final33;set final;modelo=33;

%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
leafsize=25,
iteraciones=200,shrink=0.001,maxbranch=4,maxdepth=8,mincatsize=5,minob
s=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final34;set final;modelo=34;

%cruzadatreeboostbin(archivo=churn,
vardepen=Churn,
conti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
```

```
categor=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
leafsize=20,
iteraciones=200,shrink=0.01,maxbranch=4,maxdepth=8,mincatsize=10,minob
s=20,
ngrupos=10,sinicio=13335,sfinal=13345,objetivo=tasafallos);
data final35;set final;modelo=35;

title j=c 'GBM1 - GBM2 - GBM3 - GBM4 - GBM5 - GBM6';
data union;set final30 final31 final32 final33 final34 final35;
proc boxplot data=union;plot media*modelo;run;
/* El modelo34 tiene un shrink demasiado bajo y por eso no es competitivo
con los demás*/

title j=c 'GBM1 - GBM2 - GBM3 - GBM4 - GBM6';
data union;set final30 final31 final32 final33 final35;
proc boxplot data=union;plot media*modelo;run;
/* El modelo33 es el mejor de los GBM utilizados*/

/* SUPPORT VECTOR MACHINE */

%cruzadaSVMbin
(archivo=churn,
vardepen=Churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
listclass=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,kernel=rbf k_par=10,c=10);
data final40;set final;modelo='SVM-RBF';
/* no converge */


%cruzadaSVMbin
(archivo=churn,
vardepen=Churn,
listconti=Age  Avg_Monthly_GB_Download  Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
listclass=Contract  Device_Protection_Plan  Gender_Female  Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
```

```
ngrupos=10,sinicio=12345,sfinal=12365,kernel=polynom k_par=2,c=10);
data final41;set final;modelo='SVM-poly';
/* no converge */

%cruzadaSVMbin
(archivo=churn,
vardepen=Churn,
listconti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
listclass=Contract Device_Protection_Plan Gender_Female Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,kernel=polynom k_par=3,c=10);
data final42;set final;modelo='SVM1';
/* no converge */

%cruzadaSVMbin
(archivo=churn,
vardepen=Churn,
listconti=Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
listclass=Contract Device_Protection_Plan Gender_Female Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,sinicio=12345,sfinal=12365,kernel=lineal, c=10);
data final43;set final;modelo='SVM_LIN';


proc boxplot data=final43;plot media*modelo;run;
options notes;

/*-----------------------------------------------------------------------
-----------------------------------------------------*/

/* ENSAMBLADO */
/* Ensamblado con SVM */

%cruzadastackcon(archivo=churn,
vardepen=Churn,
listconti= Age Avg_Monthly_GB_Download Avg_Monthly_Long_Distance_Charge
CLTV
Monthly_Charge      Population      Tenure_in_Months      Total_Charges
Total_Extra_Data_Charges
Total_Long_Distance_Charges Total_Refunds,
listclass=Contract Device_Protection_Plan Gender_Female Internet_Type
Married
Multiple_Lines    Number_of_Dependents    Number_of_Referrals    Offer
Online_Backup
```

```
Online_Security Paperless_Billing Partner Payment_Method Phone_Service
Premium_Tech_Support Streaming_Movies Streaming_Music Streaming_TV
Unlimited_Data,
ngrupos=10,seminicio=22345,semifinal=22355,
nodos=7,algo=bprop   mom=0.3   learn=0.2,rediter=100,/*parámetros   red
BPROP*/
maxtrees=150,vars_to_try=8,trainfraction=0.8,leafsize=25,maxdepth=8,/*
random forest */
bleafsize=25,iterations=200,bmaxbranch=4,bmaxdepth=8,shrinkage=0.01/* g
boosting*/,
kernel=lineal,degree=2,k_par=1,c=10); /* SVM Lineal*/


/*PREPARACION GRAFICO Y ETIQUETAS */

data cajas;
array ase{23};
set final;
do i=1 to 23;
modelo=i;
error=ase{i};
output;
end;
run;

proc sort data=cajas;by modelo;
data eti;length eti $ 13;
input modelo eti $;
cards;
1 RED
2 LOG
3 RFOR
4 BOOST
5 SVM
6 RLOG
7 REDFOR
8 REDBOO
9 LRFOR
10 LBOOST
11 RFORBOO
12 R-L-RFOR
13 R-L-BOO
14 R-RF-BOO
15 L-RF-BOO
16 R-L-RF-BOO
17 15ponde
18 R-SVM
19 RF-SVM
20 L-SVM
21 BOO-SVM
22 SVMLRF
23 RLRFBSVM
;
data cajas2;merge cajas eti;by modelo;
title1
h=2 box=1 j=c c=red 'Churn Prediction (con SVM)' j=c ;

options font="Courier New" bold 10;
run;goptions htext=6pt;

ods graphics off;
```

```
proc boxplot data=cajas2;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
;run;

/* 1ra selección de mejores modelos */
/* Se quitan los peores modelos*/
data cajas3;
set cajas2;
   if eti = 'RFOR' then delete;
   if eti = 'SVM' then delete;
   if eti = 'RF-SVM' then delete;
   if eti = 'SVMLRF' then delete;
   if eti = '15ponde' then delete;
   if eti = 'RFORBOO' then delete;
   if eti = 'LRFOR' then delete;
   if eti = 'REDFOR' then delete;
run;

ods graphics off;

proc boxplot data=cajas3;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
run;


/* 2da selección de mejores modelos */
/* Se quitan los peores modelos*/
data cajas4;
set cajas3;
   if eti = 'LOG' then delete;
   if eti = 'RLOG' then delete;
   if eti = 'R-L-RFOR' then delete;
   if eti = 'R-RF-BOO' then delete;
   if eti = 'L-RF-BOO' then delete;
   if eti = 'R-SVM' then delete;
   if eti = 'L-SVM' then delete;
   if eti = 'BOO-SVM' then delete;
   if eti = 'RLRFBSVM' then delete;
RUN;

ods graphics off;

proc boxplot data=cajas4;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
;run;

/* Mejores modelos:
   RED
   BOOST
   REDBOO (Mejor modelo) --> Por debajo de 0,15 de tasa de fallo
   LBOOST
   R-L-BOO
   R-L-RF-BOO (la inclusión de Random Forest empeora este ensamblado)
*/
```

15 Anexo V – Código utilizado en R

15.1 Funciones

```
library(plyr)
detach(package:plyr)
library(dummies)
library(MASS)
library(reshape)
library(caret)
library(dplyr)
library(pROC)
library(adabag)


# ********************************
# CRUZADA LOGISTICA
# ********************************

cruzadalogistica <- function(data=data,vardep=NULL,
 listconti=NULL,listclass=NULL,grupos=4,sinicio=1234,repe=5)
{

 if  (listclass !=c(""))
 {
  for (i in 1:dim(array(listclass))) {
   numindi<-which(names(data)==listclass[[i]])
   data[,numindi]<-as.character(data[,numindi])
   data[,numindi]<-as.factor(data[,numindi])
  }
 }

 data[,vardep]<-as.factor(data[,vardep])

 # Creo la formula para la logistica

 if  (listclass!=c(""))
 {
  koko<-c(listconti,listclass)
 }  else   {
  koko<-c(listconti)
 }

 modelo<-paste(koko,sep="",collapse="+")
 formu<-formula(paste(vardep,"~",modelo,sep=""))

 formu
 # Preparo caret

 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)

 # Aplico caret y construyo modelo

 regresion <- train(formu,data=data,
  trControl=control,method="glm",family = binomial(link="logit"))
 preditest<-regresion$pred
```

```
 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }

 # Aplicamos función sobre cada Repetición

 medias<-preditest %>%
  group_by(Rep) %>%
  summarize(tasa=1-tasafallos(pred,obs))

 # CalculamoS AUC  por cada Repetición de cv
 # Definimnos función

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

return(list(medias,preditest))

 }


# ********************************
# CRUZADA avNNet
# **************

cruzadaavnnetbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",grupos=4,sinicio=1234,repe=5,
  size=c(5),decay=c(0.01),repeticiones=5,itera=100,trace=FALSE)
 {

  # Preparación del archivo

  # b)pasar las categóricas a dummies

  if  (listclass!=c(""))
```

```
{
 databis<-data[,c(vardep,listconti,listclass)]
 databis<- dummy.data.frame(databis, listclass, sep = ".")
} else  {
 databis<-data[,c(vardep,listconti)]
}

# c)estandarizar las variables continuas

# Calculo medias y dtipica de datos y estandarizo (solo las continuas)

means <-apply(databis[,listconti],2,mean)
sds<-sapply(databis[,listconti],sd)

# Estandarizo solo las continuas y uno con las categoricas

datacon<-scale(databis[,listconti], center = means, scale = sds)
numerocont<-which(colnames(databis)%in%listconti)
databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

databis[,vardep]<-as.factor(databis[,vardep])

formu<-formula(paste(vardep,"~.",sep=""))

# Preparo caret

set.seed(sinicio)
control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
 savePredictions = "all",classProbs=TRUE)

# Aplico caret y construyo modelo

avnnetgrid <-  expand.grid(size=size,decay=decay,bag=FALSE)

avnnet<- train(formu,data=databis,
 method="avNNet",linout = FALSE,maxit=itera,repeats=repeticiones,
 trControl=control,tuneGrid=avnnetgrid,trace=trace)

print(avnnet$results)

preditest<-avnnet$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}

# Aplicamos función sobre cada Repetición

medias<-preditest %>%
 group_by(Rep) %>%
```

```
  summarize(tasa=1-tasafallos(pred,obs))

 # CalculamoS AUC  por cada Repetición de cv
 # Definimnos función

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

}


# ********************************
# CRUZADA Random Forest
# ******************************


cruzadarfbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,nodesize=20,
  mtry=2,ntree=50,replace=TRUE,sampsize=1)
 {


  # Preparación del archivo

  # b)pasar las categóricas a dummies

  if  (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else   {
   databis<-data[,c(vardep,listconti)]
  }

  # c)estandarizar las variables continuas

  # Calculo medias y dtipica de datos y estandarizo (solo las continuas)

  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)
```

```
 # Estandarizo solo las continuas y uno con las categoricas

 datacon<-scale(databis[,listconti], center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
 databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

 databis[,vardep]<-as.factor(databis[,vardep])

 formu<-formula(paste("factor(",vardep,")~.",sep=""))

 # Preparo caret

 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)

 # Aplico caret y construyo modelo

 rfgrid <-expand.grid(mtry=mtry)

    if  (sampsize==1)
 {
  rf<- train(formu,data=databis,
  method="rf",trControl=control,
  tuneGrid=rfgrid,nodesize=nodesize,replace=replace,ntree=ntree)
 }

else  if  (sampsize!=1)
 {
   rf<- train(formu,data=databis,
  method="rf",trControl=control,
  tuneGrid=rfgrid,nodesize=nodesize,replace=replace,sampsize=sampsize,
  ntree=ntree)
 }

 print(rf$results)

 preditest<-rf$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }

 # Aplicamos función sobre cada Repetición

 medias<-preditest %>%
  group_by(Rep) %>%
  summarize(tasa=1-tasafallos(pred,obs))
```

```
 # CalculamoS AUC  por cada Repetición de cv
 # Definimnos función

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

 }

# ************************************************************
# gbm : parámetros

#
#    Number of Boosting Iterations (n.trees, numeric)
#    Max Tree Depth (max.depth, numeric)
#    Shrinkage (shrinkage, numeric)
#    Min. Terminal Node Size (n.minobsinnode, numeric)
#
# ************************************************************

cruzadagbmbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  n.minobsinnode=20,shrinkage=0.1,n.trees=100,interaction.depth=2)
 {

  # Preparación del archivo

  # b)pasar las categóricas a dummies

  if  (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }

  # c)estandarizar las variables continuas

  # Calculo medias y dtipica de datos y estandarizo (solo las continuas)
```

```
means <-apply(databis[,listconti],2,mean)
sds<-sapply(databis[,listconti],sd)

# Estandarizo solo las continuas y uno con las categoricas

datacon<-scale(databis[,listconti], center = means, scale = sds)
numerocont<-which(colnames(databis)%in%listconti)
databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

databis[,vardep]<-as.factor(databis[,vardep])

formu<-formula(paste("factor(",vardep,")~.",sep=""))

# Preparo caret

set.seed(sinicio)
control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
 savePredictions = "all",classProbs=TRUE)

# Aplico caret y construyo modelo


 gbmgrid <-expand.grid(n.minobsinnode=n.minobsinnode,
  shrinkage=shrinkage,n.trees=n.trees,
  interaction.depth=interaction.depth)

gbm<- train(formu,data=databis,
 method="gbm",trControl=control,
 tuneGrid=gbmgrid,distribution="bernoulli",verbose=FALSE)

print(gbm$results)

preditest<-gbm$pred


preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}

# Aplicamos función sobre cada Repetición

medias<-preditest %>%
 group_by(Rep) %>%
 summarize(tasa=1-tasafallos(pred,obs))

# CalculamoS AUC  por cada Repetición de cv
# Definimnos función

auc<-function(x,y) {
 curvaroc<-roc(response=x,predictor=y)
```

```
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

}


# ************************************************************
# xgboost: parámetros

  # nrounds (# Boosting Iterations)
   # max_depth (Max Tree Depth)
   # eta (Shrinkage)
   # gamma (Minimum Loss Reduction)
   # colsample_bytree (Subsample Ratio of Columns)
   # min_child_weight (Minimum Sum of Instance Weight)
   # subsample (Subsample Percentage)
#
# ************************************************************

cruzadaxgbmbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=20,eta=0.1,nrounds=100,max_depth=2,
  gamma=0,colsample_bytree=1,subsample=1,alpha=0,lambda=0,lambda_bias=0)
 {

  # Preparación del archivo

  # b)pasar las categóricas a dummies

  if  (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }

  # c)estandarizar las variables continuas

  # Calculo medias y dtipica de datos y estandarizo (solo las continuas)

  means <-apply(databis[,listconti],2,mean)
```

```
sds<-sapply(databis[,listconti],sd)

# Estandarizo solo las continuas y uno con las categoricas

datacon<-scale(databis[,listconti], center = means, scale = sds)
numerocont<-which(colnames(databis)%in%listconti)
databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

databis[,vardep]<-as.factor(databis[,vardep])

formu<-formula(paste("factor(",vardep,")~.",sep=""))

# Preparo caret

set.seed(sinicio)
control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
 savePredictions = "all",classProbs=TRUE)

# Aplico caret y construyo modelo

xgbmgrid <-expand.grid( min_child_weight=min_child_weight,
eta=eta,nrounds=nrounds,max_depth=max_depth,
gamma=gamma,colsample_bytree=colsample_bytree,subsample=subsample)

xgbm<- train(formu,data=databis,
 method="xgbTree",trControl=control,
 tuneGrid=xgbmgrid,objective = "binary:logistic",verbose=FALSE,
 alpha=alpha,lambda=lambda,lambda_bias=lambda_bias)

print(xgbm$results)

preditest<-xgbm$pred


preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}

# Aplicamos función sobre cada Repetición

medias<-preditest %>%
 group_by(Rep) %>%
 summarize(tasa=1-tasafallos(pred,obs))

# CalculamoS AUC  por cada Repetición de cv
# Definimnos función

auc<-function(x,y) {
 curvaroc<-roc(response=x,predictor=y)
 auc<-curvaroc$auc
```

```
 return(auc)
}

# Aplicamos función sobre cada Repetición

mediasbis<-preditest %>%
 group_by(Rep) %>%
 summarize(auc=auc(obs,Yes))

# Unimos la info de auc y de tasafallos

medias$auc<-mediasbis$auc

 return(list(medias,preditest))

}



# ************************************************************
# svmLinear: parámetros

# Cost (C, numeric)

# ************************************************************


cruzadaSVMbin<-
function(data=data,vardep="vardep",
 listconti="listconti",listclass="listclass",
 grupos=4,sinicio=1234,repe=5,
 C=1,replace=TRUE)
{

 # Preparación del archivo

 # b)pasar las categóricas a dummies

 if  (listclass!=c(""))
 {
  databis<-data[,c(vardep,listconti,listclass)]
  databis<- dummy.data.frame(databis, listclass, sep = ".")
 }  else   {
  databis<-data[,c(vardep,listconti)]
 }

 # c)estandarizar las variables continuas

 # Calculo medias y dtipica de datos y estandarizo (solo las continuas)

 means <-apply(databis[,listconti],2,mean)
 sds<-sapply(databis[,listconti],sd)

 # Estandarizo solo las continuas y uno con las categoricas

 datacon<-scale(databis[,listconti], center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
```

```
databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

databis[,vardep]<-as.factor(databis[,vardep])

formu<-formula(paste("factor(",vardep,")~.",sep=""))

# Preparo caret

set.seed(sinicio)
control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
 savePredictions = "all",classProbs=TRUE)

# Aplico caret y construyo modelo

SVMgrid <-expand.grid(C=C)

SVM<- train(formu,data=databis,
 method="svmLinear",trControl=control,
 tuneGrid=SVMgrid,replace=replace)

print(SVM$results)

preditest<-SVM$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}

# Aplicamos función sobre cada Repetición

medias<-preditest %>%
 group_by(Rep) %>%
 summarize(tasa=1-tasafallos(pred,obs))

# CalculamoS AUC  por cada Repetición de cv
# Definimnos función

auc<-function(x,y) {
 curvaroc<-roc(response=x,predictor=y)
 auc<-curvaroc$auc
 return(auc)
}

# Aplicamos función sobre cada Repetición

mediasbis<-preditest %>%
 group_by(Rep) %>%
 summarize(auc=auc(obs,Yes))

# Unimos la info de auc y de tasafallos
```

```
 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

}


cruzadaSVMbinPoly<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  C=1,degree=2,scale=1)
 {

  # Preparación del archivo

  # b)pasar las categóricas a dummies

  if  (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else   {
   databis<-data[,c(vardep,listconti)]
  }

  # c)estandarizar las variables continuas

  # Calculo medias y dtipica de datos y estandarizo (solo las continuas)

  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)

  # Estandarizo solo las continuas y uno con las categoricas

  datacon<-scale(databis[,listconti], center = means, scale = sds)
  numerocont<-which(colnames(databis)%in%listconti)
  databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

  databis[,vardep]<-as.factor(databis[,vardep])

  formu<-formula(paste("factor(",vardep,")~.",sep=""))

  # Preparo caret

  set.seed(sinicio)
  control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
   savePredictions = "all",classProbs=TRUE)

  # Aplico caret y construyo modelo

  SVMgrid <-expand.grid(C=C,degree=degree,scale=scale)

  SVM<- train(formu,data=databis,
   method="svmPoly",trControl=control,
   tuneGrid=SVMgrid,replace=replace)
```

```
 print(SVM$results)

 preditest<-SVM$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }

 # Aplicamos función sobre cada Repetición

 medias<-preditest %>%
  group_by(Rep) %>%
  summarize(tasa=1-tasafallos(pred,obs))

 # CalculamoS AUC  por cada Repetición de cv
 # Definimnos función

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

}


# ************************************************************
# svmRadial: parámetros

   # Sigma (sigma, numeric)
   # Cost (C, numeric)

# ************************************************************



cruzadaSVMbinRBF<-
```

```
function(data=data,vardep="vardep",
 listconti="listconti",listclass="listclass",
 grupos=4,sinicio=1234,repe=5,
 C=1,sigma=1)
{

 # Preparación del archivo

 # b)pasar las categóricas a dummies

 if  (listclass!=c(""))
 {
  databis<-data[,c(vardep,listconti,listclass)]
  databis<- dummy.data.frame(databis, listclass, sep = ".")
 } else   {
  databis<-data[,c(vardep,listconti)]
 }

 # c)estandarizar las variables continuas

 # Calculo medias y dtipica de datos y estandarizo (solo las continuas)

 means <-apply(databis[,listconti],2,mean)
 sds<-sapply(databis[,listconti],sd)

 # Estandarizo solo las continuas y uno con las categoricas

 datacon<-scale(databis[,listconti], center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
 databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

 databis[,vardep]<-as.factor(databis[,vardep])

 formu<-formula(paste("factor(",vardep,")~.",sep=""))

 # Preparo caret

 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)

 # Aplico caret y construyo modelo

 SVMgrid <-expand.grid(C=C,sigma=sigma)

 SVM<- train(formu,data=databis,
  method="svmRadial",trControl=control,
  tuneGrid=SVMgrid,replace=replace)

 print(SVM$results)

 preditest<-SVM$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL
```

```
 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }

 # Aplicamos función sobre cada Repetición

 medias<-preditest %>%
  group_by(Rep) %>%
  summarize(tasa=1-tasafallos(pred,obs))

 # CalculamoS AUC  por cada Repetición de cv
 # Definimnos función

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 # Aplicamos función sobre cada Repetición

 mediasbis<-preditest %>%
  group_by(Rep) %>%
  summarize(auc=auc(obs,Yes))

 # Unimos la info de auc y de tasafallos

 medias$auc<-mediasbis$auc

  return(list(medias,preditest))

 }
```

15.2 Desarrollo

```
# Librerias
library(openxlsx)
library(dummies)
library(corrplot)
library(MASS)
library(caret)
library(randomForest)
library(reshape)
library(dplyr)
library(pROC)

source("D:/Fede/Google Drive/Master Minería de Datos e Inteligencia de Negocios/Materias/Tecnicas de Machine Learning/Trabajos/Clasificación/R/Ensamblado/cruzadas ensamblado binaria fuente.R")

# Se carga el archivo
path='D:/Fede/Google Drive/Master Minería de Datos e Inteligencia de Negocios/Materias/Tecnicas de Machine Learning/Trabajos/Clasificación/'
data <- read.xlsx(paste0(path,'churn_clean.xlsx'),colNames=TRUE)
# Frecuencia de variable objetivo
```

```
table(data$Churn)
# 0    1
# 5174 1869

# Observo el nombre de las variables
dput(names(data))

c("CustomerID", "Age", "Population", "Offer", "Contract", "CLTV",
  "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge", "Internet_Type",
  "Avg_Monthly_GB_Download", "Payment_Method", "Monthly_Charge",
  "Total_Charges", "Total_Extra_Data_Charges", "Total_Long_Distance_Charges",
  "Satisfaction_Score", "Churn", "Device_Protection_Plan", "Gender.Female",
  "Married", "Multiple_Lines", "Number_of_Dependents", "Number_of_Referrals",
  "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
  "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
  "Streaming_Music", "Streaming_TV", "Unlimited_Data", "Total_Refunds"
)
# Configuro el tipo de las variables
listclassall<-c("Offer", "Contract", "Internet_Type", "Payment_Method",
             "Number_of_Dependents", "Number_of_Referrals",
             "Device_Protection_Plan", "Gender_Female",  "Married",
             "Multiple_Lines", "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
             "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
             "Streaming_Music", "Streaming_TV", "Unlimited_Data" )
listclass<-c( "Offer", "Contract", "Internet_Type", "Payment_Method",
            "Number_of_Dependents", "Number_of_Referrals" )
listfactor<-c( "Device_Protection_Plan", "Gender_Female",  "Married",
             "Multiple_Lines", "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
             "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
             "Streaming_Music", "Streaming_TV", "Unlimited_Data" )
listconti<-c("Age", "Population", "CLTV", "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge",
           "Avg_Monthly_GB_Download",              "Monthly_Charge",                          "Total_Charges",
"Total_Extra_Data_Charges",
           "Total_Long_Distance_Charges", "Total_Refunds" )
vardep<-"Churn"
data$Churn<-as.factor(data$Churn) # Convierto la variable objetivo en factor
table(data$Churn)
# Revisión de valores nulos
table(is.na(data)) # No hay valores nulos

# convierto las binarias a integer
data[,listfactor]<-sapply(data[,listfactor],as.integer)

#  Orden a las columnas
dput(names(data)) # Verifico el nombre de las variables y su orden

c("CustomerID", "Age", "Population", "Offer", "Contract", "CLTV",
  "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge", "Internet_Type",
  "Avg_Monthly_GB_Download", "Payment_Method", "Monthly_Charge",
  "Total_Charges", "Total_Extra_Data_Charges", "Total_Long_Distance_Charges",
  "Churn", "Device_Protection_Plan", "Gender_Female",
  "Married", "Multiple_Lines", "Number_of_Dependents", "Number_of_Referrals",
  "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
  "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
  "Streaming_Music", "Streaming_TV", "Unlimited_Data", "Total_Refunds"
)
```

```
# Se pone la variable dependiente primero y se elimina "customerID"
var<-c("Churn", "Age", "Population", "Offer", "Contract", "CLTV",
     "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge", "Internet_Type",
     "Avg_Monthly_GB_Download", "Payment_Method", "Monthly_Charge",
     "Total_Charges", "Total_Extra_Data_Charges", "Total_Long_Distance_Charges",
     "Device_Protection_Plan", "Gender_Female",
     "Married", "Multiple_Lines", "Number_of_Dependents", "Number_of_Referrals",
     "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
     "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
     "Streaming_Music", "Streaming_TV", "Unlimited_Data", "Total_Refunds"
)

data<-data[,var]
# a)Eliminar las observaciones con missing en alguna variable
data2<-na.omit(data,(!is.na(data)))
# b)pasar las categóricas a dummies
data3<- dummy.data.frame(data2, listclass, sep = ".")
# c)estandarizar las variables continuas
# Calculo medias y dtipica de datos y estandarizo (solo las continuas)
means <-apply(data3[,listconti],2,mean)
sds<-sapply(data3[,listconti],sd)
# Estandarizo solo las continuas y uno con las categoricas
databis<-scale(data3[,listconti], center = means, scale = sds)
numerocont<-which(colnames(data3)%in%listconti)
databis<-cbind(databis,data3[,-numerocont])

# Nombres de variables
dput(names(databis))
c("Age", "Population", "CLTV", "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge",
  "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
  "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds",
  "Churn", "Offer.None", "Offer.Offer A", "Offer.Offer B", "Offer.Offer C",
  "Offer.Offer D", "Offer.Offer E", "Contract.Month-to-Month",
  "Contract.One Year", "Contract.Two Year", "Internet_Type.Cable",
  "Internet_Type.DSL", "Internet_Type.Fiber Optic", "Internet_Type.None",
  "Payment_Method.Bank Withdrawal", "Payment_Method.Credit Card",
  "Payment_Method.Mailed Check",
  "Device_Protection_Plan", "Gender_Female", "Married", "Multiple_Lines",
  "Number_of_Dependents.0", "Number_of_Dependents.1", "Number_of_Dependents.2",
  "Number_of_Dependents.3", "Number_of_Referrals.0", "Number_of_Referrals.1",
  "Number_of_Referrals.2", "Number_of_Referrals.3", "Number_of_Referrals.4",
  "Number_of_Referrals.5", "Number_of_Referrals.6", "Number_of_Referrals.7",
  "Number_of_Referrals.8", "Number_of_Referrals.9", "Number_of_Referrals.10",
  "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
  "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
  "Streaming_Music", "Streaming_TV", "Unlimited_Data")

var<-c("Churn","Age", "Population", "CLTV", "Tenure_in_Months",
"Avg_Monthly_Long_Distance_Charge",
  "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
  "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds",
  "Offer.None", "Offer.Offer A", "Offer.Offer B", "Offer.Offer C",
  "Offer.Offer D", "Offer.Offer E", "Contract.Month-to-Month",
  "Contract.One Year", "Contract.Two Year", "Internet_Type.Cable",
  "Internet_Type.DSL", "Internet_Type.Fiber Optic", "Internet_Type.None",
  "Payment_Method.Bank Withdrawal", "Payment_Method.Credit Card",
  "Payment_Method.Mailed Check",
```

```
 "Device_Protection_Plan", "Gender_Female", "Married", "Multiple_Lines",
 "Number_of_Dependents.0", "Number_of_Dependents.1", "Number_of_Dependents.2",
 "Number_of_Dependents.3", "Number_of_Referrals.0", "Number_of_Referrals.1",
 "Number_of_Referrals.2", "Number_of_Referrals.3", "Number_of_Referrals.4",
 "Number_of_Referrals.5", "Number_of_Referrals.6", "Number_of_Referrals.7",
 "Number_of_Referrals.8", "Number_of_Referrals.9", "Number_of_Referrals.10",
 "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
 "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
 "Streaming_Music", "Streaming_TV", "Unlimited_Data")

# se reordenan las columnas
databis<-databis[,var]

# Se renombran las columnas
colnames(databis)<-c("Churn","Age", "Population", "CLTV", "Tenure_in_Months",
"Avg_Monthly_Long_Distance_Charge",
            "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
            "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds",
            "Offer_None", "Offer_Offer_A", "Offer_Offer_B", "Offer_Offer_C",
            "Offer_Offer_D", "Offer_Offer_E", "Contract_Month_to_Month",
            "Contract_One_Year", "Contract_Two_Year", "Internet_Type_Cable",
            "Internet_Type_DSL", "Internet_Type_Fiber_Optic", "Internet_Type_None",
            "Payment_Method_Bank_Withdrawal", "Payment_Method_Credit_Card",
            "Payment_Method_Mailed_Check",
            "Device_Protection_Plan", "Gender_Female", "Married", "Multiple_Lines",
            "Number_of_Dependents_0", "Number_of_Dependents_1", "Number_of_Dependents_2",
            "Number_of_Dependents_3", "Number_of_Referrals_0", "Number_of_Referrals_1",
            "Number_of_Referrals_2", "Number_of_Referrals_3", "Number_of_Referrals_4",
            "Number_of_Referrals_5", "Number_of_Referrals_6", "Number_of_Referrals_7",
            "Number_of_Referrals_8", "Number_of_Referrals_9", "Number_of_Referrals_10",
            "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
            "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
            "Streaming_Music", "Streaming_TV", "Unlimited_Data")

# Se quita la variable Customer_ID, se pone primera la variable dependiente
# Se elimina una categoria de las variables categóricas (se utilizan k categorias menos 1)
var_k1<-c("Churn","Age", "Population", "CLTV", "Tenure_in_Months",
"Avg_Monthly_Long_Distance_Charge",
      "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
      "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds",
      "Offer_None", "Offer_Offer_A", "Offer_Offer_B", "Offer_Offer_C",
      "Offer_Offer_D",  "Contract_Month_to_Month",
      "Contract_One_Year",  "Internet_Type_Cable",
      "Internet_Type_DSL", "Internet_Type_Fiber_Optic",
      "Payment_Method_Bank_Withdrawal", "Payment_Method_Credit_Card",
      "Device_Protection_Plan", "Gender_Female", "Married", "Multiple_Lines",
      "Number_of_Dependents_0", "Number_of_Dependents_1", "Number_of_Dependents_2",
      "Number_of_Referrals_0", "Number_of_Referrals_1",
      "Number_of_Referrals_2", "Number_of_Referrals_3", "Number_of_Referrals_4",
      "Number_of_Referrals_5", "Number_of_Referrals_6", "Number_of_Referrals_7",
      "Number_of_Referrals_8", "Number_of_Referrals_9",
      "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
      "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
      "Streaming_Music", "Streaming_TV", "Unlimited_Data")

data1 <- databis[,var_k1]
```

```
#-------------------------------------------------------------------------------------------------------------#

##  Selección de variables  ##
data1$Churn<-as.factor(data1$Churn) # Convierto la variable objetivo en factor
full<-glm(Churn~.,data=data1,family = binomial(link="logit"))
null<-glm(Churn~1,data=data1,family = binomial(link="logit"))

# se busca el mejor set de variables con Stepwise
selec1<-stepAIC(null,scope=list(upper=full),direction="both",
            trace=FALSE,family = binomial(link="logit"))

summary(selec1) # Coeficientes y variables

formula(selec1) # variables seleccionadas

# Variables seleccionadas
varsel<-c("Churn",
"Contract_Month_to_Month","Number_of_Dependents_0","Number_of_Referrals_1",
          "Internet_Type_Fiber_Optic","Tenure_in_Months","Monthly_Charge",
          "Number_of_Referrals_0","Payment_Method_Credit_Card","Contract_One_Year",
          "Age","Phone_Service","Online_Security","Premium_Tech_Support",
          "Offer_Offer_D","Offer_Offer_A","Payment_Method_Bank_Withdrawal",
          "Paperless_Billing","Population","Number_of_Referrals_8",
          "Online_Backup","Total_Charges","Number_of_Referrals_9","Number_of_Referrals_7",
          "Number_of_Referrals_6","Married","Internet_Type_DSL","Streaming_Movies",
          "Streaming_Music","Device_Protection_Plan","Streaming_TV",
          "Internet_Type_Cable","Multiple_Lines","Number_of_Referrals_3",
          "Number_of_Referrals_5","Number_of_Referrals_2","Number_of_Referrals_4",
          "Total_Refunds")

data_sel<-data1[,varsel]

length(varsel) # usa 38 variables

#-------------------------------------------------------------------------------------------------------------#

## Regresión Logistica ##

data1$Churn<-as.factor(data1$Churn)

# todas las variables
set.seed(12345)
control<-trainControl(method = "repeatedcv",number=4,repeats=5,
                  savePredictions = "all",classProbs=TRUE)

# Aplico caret y construyo modelo

logistica <- train(Churn~.,data=data1,
                trControl=control,
                method="glm",
                family = binomial(link="logit"))

logistica$pred
logistica
# Accuracy   Kappa
# 0.8434189  0.5893285
```

```
# Variables seleccionadas
set.seed(12345)

# Aplico caret y construyo modelo
logistica_sel <- train(Churn~.,data=data_sel,
               trControl=control,
               method="glm",
               family = binomial(link="logit"))

logistica_sel$pred
logistica_sel
# Accuracy   Kappa
# 0.8441289  0.5906744

#-------------------------------------------------------------------------------------------------#
## Red Neuronal ##

set.seed(12346)
# Validación cruzada repetida
control<-trainControl(method = "repeatedcv",number=4,repeats=5,
                  savePredictions = "all",classProbs=TRUE)

#Se prueban con nodos mas cercanos adaptados a los datos
avnnetgrid <-expand.grid(size=c(3,5,7,9,11,13),decay=c(0.1,0.01,0.001),bag=FALSE)

redavnnet<- train(Churn~.,data=data1,
              method="avNNet",linout = FALSE,maxit=100,trControl=control,tuneGrid=avnnetgrid,
              repeats=5)

redavnnet

redavnnet$pred


# LOGISTICA Y AVNNET
# Se usan los datos originales limpios
data$Churn<-ifelse(data$Churn==1,"Yes","No")
table(data$Churn)

# se utilizan las variables seleccionadas
data_sel$Churn<-ifelse(data_sel$Churn==1,"Yes","No")
table(data_sel$Churn)

# grupos de Cross Validation
g=10
# Repeticiones
r=10

# Logistica
medias1<-cruzadalogistica(data=data,
                    vardep=vardep,
                    listconti=listconti,
                    listclass=listclassall, grupos=g,sinicio=1234,repe=r)

medias1$modelo="Logística"

medias1bis<-as.data.frame(medias1[1])
```

```
medias1bis$modelo<-"Logistica"
predi1<-as.data.frame(medias1[2])
predi1$logi<-predi1$Yes

# redes Neuronales
medias2<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                  size=c(3),decay=c(0.1),repeticiones=5,itera=20)

medias2$modelo="Red_3N"

medias2bis<-as.data.frame(medias2[1])
medias2bis$modelo<-"Red_3N"
predi2<-as.data.frame(medias2[2])
predi2$Red_3N<-predi2$Yes


medias3<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                  size=c(5),decay=c(0.1),repeticiones=5,itera=20)

medias3$modelo="Red_5N"

medias3bis<-as.data.frame(medias3[1])
medias3bis$modelo<-"Red_5N"
predi3<-as.data.frame(medias3[2])
predi3$Red_5N<-predi3$Yes

medias4<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                  size=c(7),decay=c(0.1),repeticiones=5,itera=20)

medias4$modelo="Red_7N"

medias4bis<-as.data.frame(medias4[1])
medias4bis$modelo<-"Red_7N"
predi4<-as.data.frame(medias4[2])
predi4$Red_7N<-predi4$Yes

medias5<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                  size=c(9),decay=c(0.1),repeticiones=5,itera=20)

medias5$modelo="Red_9N"

medias5bis<-as.data.frame(medias5[1])
medias5bis$modelo<-"Red_9N"
predi5<-as.data.frame(medias5[2])
predi5$Red_9N<-predi5$Yes
```

```
medias6<-cruzadaavnnetbin(data=data,
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                size=c(11),decay=c(0.1),repeticiones=5,itera=20)

medias6$modelo="Red_11N"

medias6bis<-as.data.frame(medias6[1])
medias6bis$modelo<-"Red_11N"
predi6<-as.data.frame(medias6[2])
predi6$Red_11N<-predi6$Yes

medias7<-cruzadaavnnetbin(data=data,
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,grupos=g,sinicio=1234,repe=r,
                size=c(13),decay=c(0.1),repeticiones=5,itera=20)

medias7$modelo="Red_13N"

medias7bis<-as.data.frame(medias7[1])
medias7bis$modelo<-"Red_13N"
predi7<-as.data.frame(medias7[2])
predi7$Red_11N<-predi7$Yes

# se unen los modelos
union1<-rbind(medias1,medias2,medias3,medias4,medias5,medias6,medias7)

union1<-rbind(medias1bis,medias2bis,medias3bis,medias4bis,medias5bis,medias6bis,medias7bis)

# Se generan los graficos de Tasa de Fallos (MISC) y área bajo la curva roc (AUC)
par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")

# se unen los mejores modelos
union1<-rbind(medias1,medias6,medias7)

# Se generan los graficos de Tasa de Fallos (MISC) y área bajo la curva roc (AUC)
par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")

## La regresión Logística es la mejor
# medias1
# La red de 13 nodos no es mejor que la de 11 nodos

# usar logistica y Red 11 nodos

## LOGISTICA Y RED CON VARIABLES SELECCIONADAS ##
# Se usan los datos originales limpios con variables seleccionadas

listcontisel<-c("Tenure_in_Months","Monthly_Charge","Age","Population","Total_Refunds")
```

```
listclasssel<-c("Contract",
"Number_of_Dependents","Number_of_Referrals","Internet_Type","Payment_Method",
          "Phone_Service","Online_Security","Premium_Tech_Support", "Offer",
          "Married","Internet_Type","Streaming_Movies", "Streaming_Music",
               "Device_Protection_Plan","Streaming_TV","Multiple_Lines")

# Logistica
medias11<-cruzadalogistica(data=data,
                  vardep=vardep,
                  listconti=listcontisel,
                  listclass=listclasssel, grupos=g,sinicio=1234,repe=r)

medias11$modelo="Log_Sel"

medias11bis<-as.data.frame(medias11[1])
medias11bis$modelo<-"Log_Sel"
predi11<-as.data.frame(medias11[2])
predi11$Log_Sel<-predi11$Yes

# redes Neuronales
medias21<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listcontisel,
                  listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                  size=c(3),decay=c(0.1),repeticiones=5,itera=20)

medias21$modelo="Red_3N_Sel"

medias21bis<-as.data.frame(medias21[1])
medias21bis$modelo<-"Red_3N_Sel"
predi21<-as.data.frame(medias21[2])
predi21$Red_3N_Sel<-predi21$Yes


medias22<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listcontisel,
                  listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                  size=c(5),decay=c(0.1),repeticiones=5,itera=20)

medias22$modelo="Red_5N_Sel"

medias22bis<-as.data.frame(medias22[1])
medias22bis$modelo<-"Red_5N_Sel"
predi22<-as.data.frame(medias22[2])
predi22$Red_5N_Sel<-predi22$Yes

medias23<-cruzadaavnnetbin(data=data,
                  vardep=vardep,
                  listconti=listcontisel,
                  listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                  size=c(7),decay=c(0.1),repeticiones=5,itera=20)

medias23$modelo="Red_7N_Sel"

medias23bis<-as.data.frame(medias23[1])
medias23bis$modelo<-"Red_7N_Sel"
```

```
predi23<-as.data.frame(medias23[2])
predi23$Red_7N_Sel<-predi23$Yes

medias24<-cruzadaavnnetbin(data=data,
                vardep=vardep,
                listconti=listcontisel,
                listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                size=c(9),decay=c(0.1),repeticiones=5,itera=20)

medias24$modelo="Red_9N_Sel"

medias24bis<-as.data.frame(medias24[1])
medias24bis$modelo<-"Red_9N_Sel"
predi24<-as.data.frame(medias24[2])
predi24$Red_9N_Sel<-predi24$Yes

medias25<-cruzadaavnnetbin(data=data,
                vardep=vardep,
                listconti=listcontisel,
                listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                size=c(11),decay=c(0.1),repeticiones=5,itera=20)

medias25$modelo="Red_11N_Sel"

medias25bis<-as.data.frame(medias25[1])
medias25bis$modelo<-"Red_11N_Sel"
predi25<-as.data.frame(medias25[2])
predi25$Red_11N_Sel<-predi25$Yes

medias26<-cruzadaavnnetbin(data=data,
                vardep=vardep,
                listconti=listcontisel,
                listclass=listclasssel,grupos=g,sinicio=1234,repe=r,
                size=c(13),decay=c(0.1),repeticiones=5,itera=20)

medias26$modelo="Red_13N_Sel"

medias26bis<-as.data.frame(medias26[1])
medias26bis$modelo<-"Red_13N_Sel"
predi26<-as.data.frame(medias26[2])
predi26$Red_13N_Sel<-predi26$Yes

# se unen los modelos
union101<-rbind(medias11,medias21,medias22,medias23,medias24,medias25,medias26)

union101<-rbind(medias1bis,medias11bis,
          medias2bis,medias3bis,medias4bis,medias5bis,medias6bis,medias7bis,
          medias21bis,medias22bis,medias23bis,medias24bis,medias25bis,medias26bis)

# Se generan los graficos de Tasa de Fallos (MISC) y área bajo la curva roc (AUC)
par(cex.axis=1.2)
boxplot(data=union101,tasa~modelo,main="TASA FALLOS")
boxplot(data=union101,auc~modelo,main="AUC")

# se unen los mejores modelos
union101<-rbind(medias11,medias25,medias26)
```

```
# Se generan los graficos de Tasa de Fallos (MISC) y área bajo la curva roc (AUC)
par(cex.axis=1.2)
boxplot(data=union101,tasa~modelo,main="TASA FALLOS")
boxplot(data=union101,auc~modelo,main="AUC")


#-------------------------------------------------------------------------------------------------------#

# Random Forest #

set.seed(12345)
rfgrid<-expand.grid(mtry=c(3,9,15,21,25,31,37,43,49,52,56)) # Variables candidatas para cada nodo

# databis$Churn<-ifelse(databis$Churn==1,"Yes","No")

control<-trainControl(method = "cv",number=10,savePredictions = "all",
                      classProbs=TRUE) # isla variable dependiente es binaria es importante poner
classProbs=TRUE

rf<- train(factor(Churn)~.,data=databis,
           method="rf",trControl=control,tuneGrid=rfgrid,
           linout = FALSE,ntree=1000,sampsize=200,nodesize=10,replace=TRUE,
           importance=TRUE)

rf
# Recomienda usar un mtry = 49


rf<- train(factor(Churn)~.,data=databis,
           method="rf",trControl=control,tuneGrid=rfgrid,
           linout = FALSE,ntree=300,nodesize=10,replace=TRUE,
           importance=TRUE)

rf
# Recomienda usar un mtry = 25



# IMPORTANCIA DE VARIABLES #

final<-rf$finalModel

tabla<-as.data.frame(importance(final))
tabla<-tabla[order(-tabla$MeanDecreaseAccuracy),]
tabla
# MeanDecreaseAccuracy indica lo que baja la precision (tasa de aciertos)
# en caso de no contar con esa variable en el modelo

barplot(tabla$MeanDecreaseAccuracy,names.arg=rownames(tabla))

# PARA PLOTEAR EL ERROR OOB (Out of Bag o fuera del saco) A MEDIDA QUE AVANZAN LAS ITERACIONES
# SE USA DIRECTAMENTE EL PAQUETE randomForest
# Se calcula sobre las observaciones que no caen en el sorteo

set.seed(12345)

rfbis<-randomForest(factor(Churn)~.,data=databis,
```

```
            mtry=49,ntree=5000,sampsize=300,nodesize=10,replace=TRUE)

rfbis<-randomForest(factor(Churn)~.,data=databis,
            mtry=25,ntree=3000,nodesize=10,replace=TRUE)

plot(rfbis$err.rate[,1])


# Se prueban RF con las mejores variables

listiconti<- c("Number_of_Referrals_1", "Age","Number_of_Dependents_0",
          "Contract_Month_to_Month", "Monthly_Charge",
          "Total_Charges","Tenure_in_Months","Total_Long_Distance_Charges",
          "Avg_Monthly_GB_Download",
          "Internet_Type_Fiber_Optic","Avg_Monthly_Long_Distance_Charge",
          "Contract_Two_Year","Number_of_Referrals_0","Number_of_Referrals_9",
          "Premium_Tech_Support","Online_Security","Streaming_Music",
          "Married","Number_of_Referrals_8")

paste(listiconti, collapse="+")

set.seed(12345)
rfgrid<-expand.grid(mtry=c(3,5,7,9,11,13,15,17,19)) # Variables candidatas para cada nodo

control<-trainControl(method = "cv",number=10,savePredictions = "all",
              classProbs=TRUE)


rf<-
train(factor(Churn)~Number_of_Referrals_1+Age+Number_of_Dependents_0+Contract_Month_to_Mo
nth+
         Monthly_Charge+Total_Charges+Total_Charges+Tenure_in_Months+
         Total_Long_Distance_Charges+Avg_Monthly_GB_Download+Avg_Monthly_GB_Download+
        Internet_Type_Fiber_Optic+Avg_Monthly_Long_Distance_Charge+Contract_Two_Year+
         Number_of_Referrals_0+Number_of_Referrals_9+Premium_Tech_Support+Online_Security+
         Streaming_Music+Married+Number_of_Referrals_8,
        data=databis,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=300,nodesize=10,replace=TRUE,
        importance=TRUE)
rf
# Recomienda mtry=3 --> es decir que en cada nodo escoge 3 variables posibles a utilizar

rf2<-
train(factor(Churn)~Number_of_Referrals_1+Age+Number_of_Dependents_0+Contract_Month_to_Mo
nth+
         Monthly_Charge+Total_Charges+Total_Charges+Tenure_in_Months+
         Total_Long_Distance_Charges+Avg_Monthly_GB_Download+Avg_Monthly_GB_Download+
         Internet_Type_Fiber_Optic+Avg_Monthly_Long_Distance_Charge+Contract_Two_Year+
         Number_of_Referrals_0+Number_of_Referrals_9+Premium_Tech_Support+Online_Security+
         Streaming_Music+Married+Number_of_Referrals_8,
        data=databis,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=300,nodesize=20,replace=TRUE,
        importance=TRUE)
rf2
# recomiendo mtry=15
```

```
rf3<-
train(factor(Churn)~Number_of_Referrals_1+Age+Number_of_Dependents_0+Contract_Month_to_Mo
nth+
          Monthly_Charge+Total_Charges+Total_Charges+Tenure_in_Months+
          Total_Long_Distance_Charges+Avg_Monthly_GB_Download+Avg_Monthly_GB_Download+
          Internet_Type_Fiber_Optic+Avg_Monthly_Long_Distance_Charge+Contract_Two_Year+
          Number_of_Referrals_0+Number_of_Referrals_9+Premium_Tech_Support+Online_Security+
          Streaming_Music+Married+Number_of_Referrals_8,
        data=databis,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=300,nodesize=30,replace=TRUE,
        importance=TRUE)
rf3
# recomienda mtry=7

# Casi todos dan un Accuracy de 0,84

# TUNEADO BÁSICO DEL TAMAÑO DE MUESTRA A SORTEAR
for (muestra in seq(1000,6500,500))
{
 set.seed(12345)
 rfbis<-randomForest(factor(Churn)~.,
                data=databis,
                mtry=15,ntree=300,sampsize=muestra,nodesize=10,replace=TRUE)

 plot(rfbis$err.rate[,1],main=muestra)

}

# Ahora se comprueba con validación cruzada con caret

rfgrid<-expand.grid(mtry=c(5,10,15,20,30,40))

rf<- train(factor(Churn)~.,
        data=databis,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=200,sampsize=4000,nodesize=10,replace=TRUE)

rf

rf2<- train(factor(Churn)~.,
        data=databis,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=200,sampsize=2000,nodesize=10,replace=TRUE)

rf2


# TUNEADO BÁSICO DEL TAMAÑO DE MUESTRA A SORTEAR
for (muestra in seq(5,30,5))
{
 set.seed(12345)
 rfbis<-randomForest(factor(Churn)~.,
                data=databis,
                mtry=15,ntree=300,sampsize=2000,nodesize=muestra,replace=TRUE)
```

```
  plot(rfbis$err.rate[,1],main=muestra)

}

## ARBOL | BAGGING | RANDOM FOREST ##

# databis$Churn<-ifelse(databis$Churn=="1","Yes","No")
table(databis$Churn)

listvar<-c("Age", "Population", "CLTV", "Tenure_in_Months", "Avg_Monthly_Long_Distance_Charge",
        "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
        "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds",
        "Offer_None", "Offer_Offer_A", "Offer_Offer_B", "Offer_Offer_C",
        "Offer_Offer_D", "Offer_Offer_E", "Contract_Month_to_Month",
        "Contract_One_Year", "Contract_Two_Year", "Internet_Type_Cable",
        "Internet_Type_DSL", "Internet_Type_Fiber_Optic", "Internet_Type_None",
        "Payment_Method_Bank_Withdrawal", "Payment_Method_Credit_Card",
        "Payment_Method_Mailed_Check",
        "Device_Protection_Plan", "Gender_Female", "Married", "Multiple_Lines",
        "Number_of_Dependents_0", "Number_of_Dependents_1", "Number_of_Dependents_2",
        "Number_of_Dependents_3", "Number_of_Referrals_0", "Number_of_Referrals_1",
        "Number_of_Referrals_2", "Number_of_Referrals_3", "Number_of_Referrals_4",
        "Number_of_Referrals_5", "Number_of_Referrals_6", "Number_of_Referrals_7",
        "Number_of_Referrals_8", "Number_of_Referrals_9", "Number_of_Referrals_10",
        "Online_Backup", "Online_Security", "Paperless_Billing", "Partner",
        "Phone_Service", "Premium_Tech_Support", "Streaming_Movies",
        "Streaming_Music", "Streaming_TV", "Unlimited_Data")

medias301<-cruzadaarbolbin(data=databis,
                   vardep=vardep,
                   listconti=listvar,
                   listclass=c(""),
                   grupos=g,sinicio=1234,repe=r,
                   cp=c(0),minbucket=10)

medias301$modelo="arbol"

medias301bis<-as.data.frame(medias301[1])
medias301bis$modelo<-"arbol"
predi301<-as.data.frame(medias301[2])
predi301$arbol<-predi301$Yes

medias302<-cruzadarfbin(data=data,
                    vardep=vardep,
                    listconti=listconti,
                    listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,nodesize=10,
                  mtry=56,ntree=100,replace=TRUE)

medias302$modelo="bagging"

medias302bis<-as.data.frame(medias302[1])
medias302bis$modelo<-"bagging"
predi302<-as.data.frame(medias302[2])
predi302$bagging<-predi302$Yes

medias303<-cruzadarfbin(data=data,
```

```
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,
                grupos=g,sinicio=1234,repe=r,nodesize=10,
                mtry=20,ntree=600,replace=TRUE)

medias303$modelo="rf1"

medias303bis<-as.data.frame(medias303[1])
medias303bis$modelo<-"rf1"
predi303<-as.data.frame(medias303[2])
predi303$rf1<-predi303$Yes

medias304<-cruzadarfbin(data=data,
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,
                grupos=g,sinicio=1234,repe=r,nodesize=10,
                mtry=30,ntree=600,replace=TRUE)

medias304$modelo="rf2"

medias304bis<-as.data.frame(medias304[1])
medias304bis$modelo<-"rf2"
predi304<-as.data.frame(medias304[2])
predi304$rf2<-predi304$Yes

medias305<-cruzadarfbin(data=data,
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,
                grupos=g,sinicio=1234,repe=r,nodesize=10,
                mtry=10,ntree=600,replace=TRUE)

medias305$modelo="rf3"

medias305bis<-as.data.frame(medias305[1])
medias305bis$modelo<-"rf3"
predi305<-as.data.frame(medias305[2])
predi305$rf3<-predi305$Yes

medias306<-cruzadarfbin(data=data,
                vardep=vardep,
                listconti=listconti,
                listclass=listclassall,
                grupos=g,sinicio=1234,repe=r,nodesize=20,
                mtry=20,ntree=600,replace=TRUE)

medias306$modelo="rf4"

medias306bis<-as.data.frame(medias306[1])
medias306bis$modelo<-"rf4"
predi306<-as.data.frame(medias306[2])
predi306$rf4<-predi306$Yes

union1<-rbind(medias302,medias303,medias304,medias305)
```

```
union1<-rbind(medias302bis,medias303bis,medias304bis,medias305bis,medias306bis)

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")

#-------------------------------------------------------------------------------------------------#
##  GRADIENT BOOSTING ##

set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.1,0.05,0.03,0.01,0.001),
                n.minobsinnode=c(5,10,20,30),
                n.trees=c(100,500,1000,5000),
                interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
                 classProbs=TRUE)

# La Variable objetivo debe ser "Yes" or "No"
gbm<- train(factor(Churn)~.,data=databis,
         method="gbm",trControl=control,tuneGrid=gbmgrid,
         distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm
# Tuning parameter 'interaction.depth' was held constant at a value of 2
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were n.trees = 500, interaction.depth = 2,
# shrinkage = 0.05 and n.minobsinnode = 30.

# recomienda 500 arboles, interaction.depth = 2, shrinkage = 0.05 and n.minobsinnode = 30
# Accuracy de 0,8510578

plot(gbm)
# se debe observar si existen patrones

# Otra prueba pero con shrinkages más cercanos
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.02,0.03,0.04,0.05,0.06),
                n.minobsinnode=c(5,10,20,30),
                n.trees=c(100,500,1000,5000),
                interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
                 classProbs=TRUE)

# La Variable objetivo debe ser "Yes" or "No"
gbm<- train(factor(Churn)~.,data=databis,
         method="gbm",trControl=control,tuneGrid=gbmgrid,
         distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm
# Tuning parameter 'interaction.depth' was held constant at a value of 2
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were n.trees = 500, interaction.depth = 2,
# shrinkage = 0.04 and n.minobsinnode = 30.
```

```
# recomienda 500 arboles, interaction.depth = 2, shrinkage = 0.04 and n.minobsinnode = 30
# Accuracy de 0,8513419

plot(gbm)
# se debe observar si existen patrones

# ESTUDIO DE EARLY STOPPING
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.04),
                n.minobsinnode=c(30),
                n.trees=c(100,300,500,800,1000,1200),
                interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
                 classProbs=TRUE)

gbm<- train(factor(Churn)~.,data=databis,
         method="gbm",trControl=control,tuneGrid=gbmgrid,
         distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm
# Tuning parameter 'interaction.depth' was held constant at a value of 2
# Tuning parameter 'shrinkage' was held constant
# at a value of 0.04
# Tuning parameter 'n.minobsinnode' was held constant at a value of 30
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were n.trees = 500,
# interaction.depth = 2, shrinkage = 0.04 and n.minobsinnode = 30.

plot(gbm)

# IMPORTANCIA DE VARIABLES
par(cex=1.3)
summary(gbm)

tabla<-summary(gbm)
par(cex=1.5,las=2)
barplot(tabla$rel.inf,names.arg=row.names(tabla))


## Gradient Boosting Machine Binaria Cruzada ##

# data$Churn<-ifelse(data$Churn=="1","Yes","No")
# table(data$Churn)

medias401<-cruzadagbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,
                  n.minobsinnode=30,shrinkage=0.04,n.trees=500,interaction.depth=2)

medias401$modelo="gbm"

medias401bis<-as.data.frame(medias401[1])
```

```
medias401bis$modelo<-"gbm"
predi401<-as.data.frame(medias401[2])
predi401$gbm<-predi401$Yes

# Medias 3** son RF
union1<-rbind(medias303,medias304,medias401)

union1<-rbind(medias1bis,medias11bis,
          medias303bis,medias304bis,medias305bis,medias306bis,
          medias401bis)

union1<-rbind(medias1bis,medias11bis,medias401bis) # GBM es mejor que logística

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")


#-------------------------------------------------------------------------------------------------------------#

## EXTREME GRADIENT BOOSTING ##

set.seed(12345)

xgbmgrid<-expand.grid(
  min_child_weight=c(5,10,20),
  eta=c(0.1,0.05,0.03,0.01,0.001),
  nrounds=c(100,500,1000,5000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

control<-trainControl(method = "cv",number=4,savePredictions = "all",
                classProbs=TRUE)
inicio<-Sys.time()

# La Variable objetivo debe ser "Yes" or "No"
xgbm<- train(factor(Churn)~.,data=databis,
          method="xgbTree",trControl=control,
          tuneGrid=xgbmgrid,verbose=FALSE)

finxgboost<-Sys.time()-inicio

xgbm
# Tuning parameter 'max_depth' was held constant at a value of 6
# Tuning parameter 'gamma' was held constant at a value of 0
# Tuning parameter 'colsample_bytree' was held constant at a value of 1
# Tuning parameter 'subsample' was held constant at a value of 1
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were nrounds = 500, max_depth = 6,
# eta = 0.01, gamma = 0, colsample_bytree = 1, min_child_weight = 10 and subsample = 1.
# Accuracy 0.8477931

plot(xgbm)

# ESTUDIO DE EARLY STOPPING
xgbmgrid<-expand.grid(eta=c(0.01),
                min_child_weight=c(10),
                nrounds=c(50,100,150,200,250,300,500,800,1200,1500,2000),
```

```
            max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

set.seed(12345)
control<-trainControl(method = "cv",number=4,savePredictions = "all",
            classProbs=TRUE)

xgbm<- train(factor(Churn)~.,data=databis,
        method="xgbTree",trControl=control,
        tuneGrid=xgbmgrid,verbose=FALSE)

xgbm
# Tuning parameter 'max_depth' was held constant at a value of 6
# Tuning parameter 'eta' was held constant at a value
# held constant at a value of 1
# Tuning parameter 'min_child_weight' was held constant at a value of 10
# Tuning
# parameter 'subsample' was held constant at a value of 1
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were nrounds = 500,
# max_depth = 6, eta = 0.01, gamma = 0, colsample_bytree =  1,
# min_child_weight = 10 and subsample = 1.

plot(xgbm)
# parece que con 500 arboles funciona bien


# Se utilizan otras semillas para la CV

set.seed(12367)
control<-trainControl(method = "cv",number=4,savePredictions = "all",
            classProbs=TRUE)


xgbm1<- train(factor(Churn)~.,data=databis,
        method="xgbTree",trControl=control,
        tuneGrid=xgbmgrid,verbose=FALSE)


xgbm1
# Tuning parameter 'max_depth' was held constant at a value of 6
# Tuning parameter 'eta' was held constant at a value
# held constant at a value of 1
# Tuning parameter 'min_child_weight' was held constant at a value of 10
# Tuning
# parameter 'subsample' was held constant at a value of 1
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were nrounds = 1200, max_depth = 6,
# eta = 0.01, gamma = 0, colsample_bytree =  1,
# min_child_weight = 10 and subsample = 1.

plot(xgbm1)


# Se utilizan otras semillas para la CV

set.seed(12399)
control<-trainControl(method = "cv",number=4,savePredictions = "all",
```

```
                classProbs=TRUE)


xgbm2<- train(factor(Churn)~.,data=databis,
          method="xgbTree",trControl=control,
          tuneGrid=xgbmgrid,verbose=FALSE)


xgbm2
# Tuning parameter 'max_depth' was held constant at a value of 6
# Tuning parameter 'eta'
#
# Tuning parameter 'min_child_weight' was held constant at a value of 10
# Tuning
# parameter 'subsample' was held constant at a value of 1
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were nrounds = 800, max_depth = 6,
# eta = 0.01, gamma = 0,
# colsample_bytree = 1, min_child_weight = 10 and subsample = 1.

plot(xgbm2)


# IMPORTANCIA DE VARIABLES

varImp(xgbm)
# xgbTree variable importance
# only 20 most important variables shown (out of 56)
#                             Overall
# Contract_Month_to_Month         100.000
# Age                              23.599
# Tenure_in_Months                 20.236
# Number_of_Dependents_0           19.410
# Number_of_Referrals_1            19.138
# Monthly_Charge                   15.842
# Number_of_Referrals_0             6.984
# Payment_Method_Credit_Card        6.732
# Total_Long_Distance_Charges       5.919
# Population                        5.183
# Total_Charges                     4.952
# Internet_Type_Fiber_Optic         4.859
# CLTV                              3.894
# Contract_One_Year                 3.661
# Avg_Monthly_GB_Download           3.192
# Contract_Two_Year                 3.084
# Online_Security                   2.702
# Avg_Monthly_Long_Distance_Charge   2.502
# Paperless_Billing                 2.112
# Streaming_Music                   2.064

plot(varImp(xgbm))


# PRUEBO PARÁMETROS CON VARIABLES SELECCIONADAS

xgbmgrid<-expand.grid(
  min_child_weight=c(5,10,20,30),
```

```
  eta=c(0.1,0.05,0.03,0.01,0.001),
  nrounds=c(100,500,1000,2000,5000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

control<-trainControl(method = "cv",number=4,savePredictions = "all",
                classProbs=TRUE)

inicio<-Sys.time()

xgbm_final<- train(factor(Churn)~.,data=databis,
          method="xgbTree",trControl=control,
          tuneGrid=xgbmgrid,verbose=FALSE)

finxgboostf1 <- Sys.time() - inicio # 31 minutos

xgbm_final
# Tuning parameter 'max_depth' was held constant at a value of 6
# Tuning parameter 'gamma' was
# held constant at a value of 0
# Tuning parameter 'colsample_bytree' was held constant at a value of
# 1
# Tuning parameter 'subsample' was held constant at a value of 1
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were nrounds = 1000, max_depth = 6,
# eta = 0.01, gamma = 0, colsample_bytree = 1,
# min_child_weight = 5 and subsample = 1.

plot(xgbm_final)


# UTILIZACIÓN DE LOS PARÁMETROS  DE REGULARIZACIÓN
xgbmgrid<-expand.grid(eta=c(0.01),
                min_child_weight=c(10),
                nrounds=c(1000),
                max_depth=6,gamma=0,colsample_bytree=1,subsample=1)


df<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("lambda", "Accuracy")
colnames(df) <- x
df2<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("lambda", "Accuracy")
colnames(df2) <- x

for (lambda in seq(0,20,1))
{
  xgbm<- train(factor(Churn)~.,data=databis,
            method="xgbTree",trControl=control,
            tuneGrid=xgbmgrid,lambda=lambda,verbose=FALSE)
  cat(lambda,"\n")
  cat(xgbm$results$Accuracy,"\n")
  df[1,1]<-lambda
  df[1,2]<-xgbm$results$Accuracy
  df2<-rbind(df2,df)
}

plot(df2$lambda,df2$Accuracy)
```

```
xgbm$results$Accuracy

df<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("alpha", "Accuracy")
colnames(df) <- x
df2<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("alpha", "Accuracy")
colnames(df2) <- x

for (alpha in seq(0,20,1))
{
 xgbm<- train(factor(Churn)~.,data=databis,
          method="xgbTree",trControl=control,
          tuneGrid=xgbmgrid,alpha=alpha,verbose=FALSE)
 cat(alpha,"\n")
 cat(xgbm$results$Accuracy,"\n")
 df[1,1]<-alpha
 df[1,2]<-xgbm$results$Accuracy
 df2<-rbind(df2,df)
}

plot(df2$alpha,df2$Accuracy)

xgbm$results$Accuracy

df<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("lambda_bias", "Accuracy")
colnames(df) <- x
df2<- data.frame(matrix(ncol = 2, nrow = 0))
x <- c("lambda_bias", "Accuracy")
colnames(df2) <- x


for (lambda_bias in seq(0,20,1))
{
 xgbm<- train(factor(Churn)~.,data=databis,
          method="xgbTree",trControl=control,
          tuneGrid=xgbmgrid,lambda_bias=lambda_bias,verbose=FALSE)
 cat(lambda_bias,"\n")
 cat(xgbm$results$Accuracy,"\n")
 df[1,1]<-lambda_bias
 df[1,2]<-xgbm$results$Accuracy
 df2<-rbind(df2,df)
}

plot(df2$lambda_bias,df2$Accuracy)

xgbm$results$Accuracy


## CRUZADA EXTREMO GRADIENT BOOSTING BINARIA ##

medias501<-cruzadaxgbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
```

```
                grupos=g,sinicio=1234,repe=r,
                min_child_weight=10,eta=0.01,nrounds=800,max_depth=6,
                gamma=0,colsample_bytree=1,subsample=1,
                alpha=3,lambda=6,lambda_bias=10)

medias501$modelo="xgbm"

medias501bis<-as.data.frame(medias501[1])
medias501bis$modelo<-"xgbm"
predi501<-as.data.frame(medias501[2])
predi501$xgbm<-predi501$Yes

medias502<-cruzadaxgbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                grupos=g,sinicio=1234,repe=r,
                min_child_weight=5,eta=0.01,nrounds=1000,max_depth=6,
                gamma=0,colsample_bytree=1,subsample=1,
                alpha=6,lambda=5,lambda_bias=10)

medias502$modelo="xgbm2"

medias502bis<-as.data.frame(medias502[1])
medias502bis$modelo<-"xgbm2"
predi502<-as.data.frame(medias502[2])
predi502$xgbm2<-predi502$Yes

# Se realiza un sorteo de variables y se aumenta el número de nrounds

medias503<-cruzadaxgbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,
                  min_child_weight=10,eta=0.01,nrounds=2500,max_depth=6,
                  gamma=0,colsample_bytree=1,subsample=0.8,
                  alpha=3,lambda=6,lambda_bias=10)

medias503$modelo="xgbm3"

medias503bis<-as.data.frame(medias503[1])
medias503bis$modelo<-"xgbm3"
predi503<-as.data.frame(medias503[2])
predi503$xgbm3<-predi503$Yes

medias504<-cruzadaxgbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,
                  min_child_weight=5,eta=0.01,nrounds=4000,max_depth=6,
                  gamma=0,colsample_bytree=1,subsample=0.8,
                  alpha=6,lambda=5,lambda_bias=10)

medias504$modelo="xgbm4"
```

```
medias504bis<-as.data.frame(medias504[1])
medias504bis$modelo<-"xgbm4"
predi504<-as.data.frame(medias504[2])
predi504$xgbm4<-predi504$Yes

# Sin regularización

medias505<-cruzadaxgbmbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,
                  min_child_weight=10,eta=0.01,nrounds=800,max_depth=6,
                  gamma=0,colsample_bytree=1,subsample=1,
                  alpha=0,lambda=0,lambda_bias=0)

medias505$modelo="xgbm5"

medias505bis<-as.data.frame(medias505[1])
medias505bis$modelo<-"xgbm5"
predi505<-as.data.frame(medias505[2])
predi505$xgbm5<-predi505$Yes

# Medias 303 y medias304 son RF
# Medias 401 es GBM
union1<-
rbind(medias303,medias304,medias401,medias501,medias502,medias503,medias504,medias505)

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")


union1<-rbind(medias1bis,medias11bis,
        medias303bis,medias304bis,medias305bis,medias306bis,
        medias401bis,
        medias501bis,medias502bis,medias503bis,medias504bis,medias505bis)

union1<-rbind(medias401bis,
        medias501bis,medias502bis,medias503bis,medias504bis,medias505bis)
# union1<-rbind(medias1bis,medias11bis,medias401bis) # GBM es mejor que logística

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")


#-------------------------------------------------------------------------------------------------------#

## SUPPORT VECTOR MACHINE ##

#  SVM LINEAL: SOLO PARÁMETRO C
set.seed(12345)
SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10))

control<-trainControl(method = "cv",number=10,savePredictions = "all")
```

```
SVM0<- train(factor(Churn)~.,
        data=databis,
        method="svmLinear",trControl=control,
        tuneGrid=SVMgrid,verbose=FALSE)
# method="svmLinear" es para un SVM linear, de 1 grado de polinomio

SVM0$results
# C  Accuracy    Kappa  AccuracySD    KappaSD
# 1  0.01 0.8418307 0.5826898 0.005986878 0.01808504
# 2  0.05 0.8422570 0.5854947 0.006339270 0.01876469
# 3  0.10 0.8432507 0.5889438 0.005621542 0.01556776
# 4  0.20 0.8432505 0.5886846 0.005871212 0.01528053
# 5  0.50 0.8429663 0.5878946 0.004804144 0.01397541
# 6  1.00 0.8433926 0.5883386 0.005875798 0.01738200
# 7  2.00 0.8443864 0.5901578 0.006069697 0.01807238
# 8  5.00 0.8439595 0.5892683 0.004482483 0.01367913
# 9 10.00 0.8443857 0.5901645 0.004745303 0.01476666

plot(SVM0$results$C,SVM0$results$Accuracy,col="red",pch=16)
# parece que hay un patron, porque los resultados del SVM se encuentran muy juntos
# la mayoría de los valores se encuentran entre 0 y 2


# Rehago el grid para observar mejor el intervalo de C entre 0.1 y 1.9
SVMgrid<-expand.grid(C=c(0.1,0.3,0.5,0.7,0.9,1.1,1.3,1.5,1.7,1.9))

control<-trainControl(method = "cv",number=4,savePredictions = "all")

SVM<-train(factor(Churn)~.,
        data=databis,
        method="svmLinear",trControl=control,
        tuneGrid=SVMgrid,verbose=FALSE)

SVM$results
#    C  Accuracy    Kappa  AccuracySD    KappaSD
# 1  0.1 0.8443855 0.5912593 0.009955965 0.03014008
# 2  0.3 0.8449533 0.5928769 0.008075732 0.02609487
# 3  0.5 0.8449534 0.5929063 0.008335852 0.02630269
# 4  0.7 0.8452373 0.5932120 0.008684135 0.02745742
# 5  0.9 0.8448114 0.5919196 0.008543927 0.02642580
# 6  1.1 0.8452373 0.5931207 0.008014702 0.02504429
# 7  1.3 0.8449536 0.5922661 0.008710645 0.02607865
# 8  1.5 0.8452376 0.5930072 0.008777451 0.02624411
# 9  1.7 0.8453796 0.5937197 0.008956541 0.02695889
# 10 1.9 0.8456637 0.5943367 0.009330372 0.02756915

plot(SVM$results$C,SVM$results$Accuracy, col="red",pch=16)
# todos los valores se encuentran por encima del Accuracy 0,8430

# Cuando el SVM es lineal es un competidor directo de la regresión Logística (comparar)


#  SVM Polinomial: PARÁMETROS C, degree, scale

SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10),
              degree=c(2,3),scale=c(0.1,0.5,1,2,5))
# Degree es el grado del polinomio
```

```
control<-trainControl(method = "cv",
                      number=10,savePredictions = "all")


SVMP<- train(factor(Churn)~.,
             data=databis,
             method="svmPoly",trControl=control,
             tuneGrid=SVMgrid,verbose=FALSE)

SVMP
# Accuracy was used to select the optimal model using the largest value.
# The final values used for the model were degree = 2, scale = 0.1 and C = 0.01.

SVMP$results

# Graficos
dat<-as.data.frame(SVM$results)
library(ggplot2)

# PLOT DE DOS VARIABLES CATEGÓRICAS, UNA CONTINUA
ggplot(dat, aes(x=factor(C), y=Accuracy,
                color=factor(degree),pch=factor(scale))) +
  geom_point(position=position_dodge(width=0.5),size=3)
# se grafican 3 variables: el grado, la escala y el accuracy
# el color azul está siempre por debajo del rojo
# el factor (degree) es mejor el 2
# el scale es mejor el 0,1
# es mayor el accuracy cuanto menor es el C

# SOLO DEGREE=2
dat2<-dat[dat$degree==2,]

ggplot(dat2, aes(x=factor(C), y=Accuracy,
                 colour=factor(scale))) +
  geom_point(position=position_dodge(width=0.5),size=3)
# Con escala 0,1 siempre el accuracy es mayor al resto
# se consideraria la escala 0,1
# un valor C de 0,01 es el mejor accuracy

# SVM RBF: PARÁMETROS C, sigma
# SVM Radial Basis Function

SVMrgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10,30),
                      sigma=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10,30))

control<-trainControl(method = "cv",
                      number=4,savePredictions = "all")


SVMr<- train(factor(Churn)~.,
             data=databis,
             method="svmRadial",trControl=control,
             tuneGrid=SVMrgrid,verbose=FALSE)

SVMr
# Accuracy was used to select the optimal model using the largest value.
```

```
# The final values used for the model were sigma = 0.01 and C = 2.

SVMr$results

dat<-as.data.frame(SVMr$results)

ggplot(dat,aes(x=factor(C), y=Accuracy,
            color=factor(sigma)))+
  geom_point(position=position_dodge(width=0.5),size=3)
# Sigma tiene que ser pequeño, porque tiene alto accuracy
# entre 0,01 y 0,1
# un C igual a 0,5 y 2 parece funcionar bien
# por ejemplo: sigma=0,01; C=1


# CRUZADA SUPPORT VECTOR MACHINE BINARIA #

medias601<-cruzadaSVMbin(data=data,
                  vardep=vardep,
                  listconti=listconti,
                  listclass=listclassall,
                  grupos=g,sinicio=1234,repe=r,
                  C=0.7)

medias601$modelo="SVM"

medias601bis<-as.data.frame(medias601[1])
medias601bis$modelo<-"SVM"
predi601<-as.data.frame(medias601[2])
predi601$SVM<-predi601$Yes

medias602<-cruzadaSVMbinPoly(data=data,
                    vardep=vardep,
                    listconti=listconti,
                    listclass=listclassall,
                    grupos=g,sinicio=1234,repe=r,
                    C=0.01,degree=2,scale=0.1)

medias602$modelo="SVMPoly"

medias602bis<-as.data.frame(medias602[1])
medias602bis$modelo<-"SVMPoly"
predi602<-as.data.frame(medias602[2])
predi602$SVMPoly<-predi602$Yes


medias603<-cruzadaSVMbinRBF(data=data,
                    vardep=vardep,
                    listconti=listconti,
                    listclass=listclassall,
                    grupos=g,sinicio=1234,repe=r,
                    C=2,sigma=0.01)

medias603$modelo="SVMRBF"

medias603bis<-as.data.frame(medias603[1])
medias603bis$modelo<-"SVMRBF"
```

```
predi603<-as.data.frame(medias603[2])
predi603$SVMRBF<-predi603$Yes

union1<-rbind(medias601,medias602,medias603)

union1<-rbind(medias601bis,medias602bis,medias603bis)

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")

# GRAFICOS CON TODOS LOS MODELOS #

union<-rbind(medias1,medias2,medias3,medias4,medias5,medias6,medias7, #Logistica # Red
        medias11,medias21,medias22,medias23,medias24,medias25,medias26, #Logistica_Sel #Red_Sel
        medias301,medias302,medias303,medias304,medias305, #Bagging #RF
        medias401, #GBM
        medias501,medias502,medias503,medias504,medias505, #XGBOOST
        medias601,medias602,medias603) #SVM

union1<-rbind(medias1bis,medias2bis,medias3bis,medias4bis,medias5bis,medias6bis,medias7bis,
#Logistica # Red
        medias11bis,medias21bis,medias22bis,medias23bis,medias24bis,medias25bis,medias26bis,
#Logistica_Sel #Red_Sel
        medias302bis,medias303bis,medias304bis,medias305bis,medias306bis, #Bagging #RF
        medias401bis, #GBM
        medias501bis,medias502bis,medias503bis,medias504bis,medias505bis, #XGBOOST
        medias601bis,medias602bis,medias603bis) #SVM

# union1<-rbind(medias1bis,medias11bis,medias401bis) # GBM es mejor que logística

par(cex.axis=1.2)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
boxplot(data=union1,auc~modelo,main="AUC")

# Se ordena segun la tasa de fallos (su media)
uni<-union1
uni$modelo <- with(uni, reorder(modelo,tasa, median))
par(cex.axis=0.75,las=2)
boxplot(data=uni,tasa~modelo, col="grey",main="TASA FALLOS")

# Se ordena segun el AUC (la media)
uni<-union1
uni$modelo <- with(uni,reorder(modelo,auc, median))
par(cex.axis=0.75,las=2)
boxplot(data=uni,auc~modelo, col="grey",main="AUC")


#-----------------------------------------------------------------------------------------------------#

## ENSAMBLADO ##

# unipredi<-NULL

# CONSTRUCCIÓN DE TODOS LOS ENSAMBLADOS

unipredi<-cbind(predi1,predi2,predi3,predi4,predi5,predi6,predi7,
```

```
        predi11,predi21,predi22,predi23,predi24,predi25,predi26,
        predi301,predi302,predi303,predi304,predi305,predi306,
        predi401,
        predi501,predi502,predi503,predi504,
        predi601,predi602,predi603)

# Esto es para eliminar columnas duplicadas
unipredi<- unipredi[, !duplicated(colnames(unipredi))]

colnames(unipredi)

# Construccion de ensamblados, cambiar al gusto
# 2 modelos #
unipredi$predi700<-(unipredi$logi+unipredi$Red_11N)/2
unipredi$predi701<-(unipredi$logi+unipredi$Red_13N_Sel)/2
unipredi$predi702<-(unipredi$logi+unipredi$rf1)/2
unipredi$predi703<-(unipredi$logi+unipredi$rf2)/2
unipredi$predi704<-(unipredi$logi+unipredi$rf3)/2
unipredi$predi705<-(unipredi$logi+unipredi$gbm)/2
unipredi$predi706<-(unipredi$logi+unipredi$xgbm)/2
unipredi$predi707<-(unipredi$logi+unipredi$xgbm2)/2
unipredi$predi708<-(unipredi$logi+unipredi$xgbm3)/2
unipredi$predi709<-(unipredi$logi+unipredi$xgbm4)/2
unipredi$predi710<-(unipredi$logi+unipredi$SVM)/2
unipredi$predi711<-(unipredi$logi+unipredi$SVMPoly)/2
unipredi$predi712<-(unipredi$logi+unipredi$SVMRBF)/2
unipredi$predi713<-(unipredi$Red_13N_Sel+unipredi$rf1)/2
unipredi$predi714<-(unipredi$Red_13N_Sel+unipredi$rf2)/2
unipredi$predi715<-(unipredi$Red_13N_Sel+unipredi$rf3)/2
unipredi$predi716<-(unipredi$Red_13N_Sel+unipredi$gbm)/2
unipredi$predi717<-(unipredi$Red_13N_Sel+unipredi$xgbm)/2
unipredi$predi718<-(unipredi$Red_13N_Sel+unipredi$xgbm2)/2
unipredi$predi719<-(unipredi$Red_13N_Sel+unipredi$xgbm3)/2
unipredi$predi720<-(unipredi$Red_13N_Sel+unipredi$xgbm4)/2
unipredi$predi721<-(unipredi$Red_13N_Sel+unipredi$SVM)/2
unipredi$predi722<-(unipredi$Red_13N_Sel+unipredi$SVMPoly)/2
unipredi$predi723<-(unipredi$Red_13N_Sel+unipredi$SVMRBF)/2
unipredi$predi724<-(unipredi$rf1+unipredi$gbm)/2
unipredi$predi725<-(unipredi$rf1+unipredi$xgbm)/2
unipredi$predi726<-(unipredi$rf1+unipredi$xgbm2)/2
unipredi$predi727<-(unipredi$rf1+unipredi$xgbm3)/2
unipredi$predi728<-(unipredi$rf1+unipredi$xgbm4)/2
unipredi$predi729<-(unipredi$rf1+unipredi$SVM)/2
unipredi$predi730<-(unipredi$rf1+unipredi$SVMPoly)/2
unipredi$predi731<-(unipredi$rf1+unipredi$SVMRBF)/2
unipredi$predi732<-(unipredi$rf2+unipredi$gbm)/2
unipredi$predi733<-(unipredi$rf2+unipredi$xgbm)/2
unipredi$predi734<-(unipredi$rf2+unipredi$xgbm2)/2
unipredi$predi735<-(unipredi$rf2+unipredi$xgbm3)/2
unipredi$predi736<-(unipredi$rf2+unipredi$xgbm4)/2
unipredi$predi737<-(unipredi$rf2+unipredi$SVM)/2
unipredi$predi738<-(unipredi$rf2+unipredi$SVMPoly)/2
unipredi$predi739<-(unipredi$rf2+unipredi$SVMRBF)/2
unipredi$predi740<-(unipredi$rf3+unipredi$gbm)/2
unipredi$predi741<-(unipredi$rf3+unipredi$xgbm)/2
unipredi$predi742<-(unipredi$rf3+unipredi$xgbm2)/2
unipredi$predi743<-(unipredi$rf3+unipredi$xgbm3)/2
```

```
unipredi$predi744<-(unipredi$rf3+unipredi$xgbm4)/2
unipredi$predi745<-(unipredi$rf3+unipredi$SVM)/2
unipredi$predi746<-(unipredi$rf3+unipredi$SVMPoly)/2
unipredi$predi747<-(unipredi$rf3+unipredi$SVMRBF)/2
unipredi$predi748<-(unipredi$gbm+unipredi$xgbm)/2
unipredi$predi749<-(unipredi$gbm+unipredi$xgbm2)/2
unipredi$predi750<-(unipredi$gbm+unipredi$xgbm3)/2
unipredi$predi751<-(unipredi$gbm+unipredi$xgbm4)/2
unipredi$predi752<-(unipredi$gbm+unipredi$SVM)/2
unipredi$predi753<-(unipredi$gbm+unipredi$SVMPoly)/2
unipredi$predi754<-(unipredi$gbm+unipredi$SVMRBF)/2

# 3 modelos #
unipredi$predi755<-(unipredi$logi+unipredi$Red_11N+unipredi$rf1)/3
unipredi$predi756<-(unipredi$logi+unipredi$Red_11N+unipredi$rf2)/3
unipredi$predi757<-(unipredi$logi+unipredi$Red_11N+unipredi$rf3)/3
unipredi$predi758<-(unipredi$logi+unipredi$Red_11N+unipredi$gbm)/3
unipredi$predi759<-(unipredi$logi+unipredi$Red_11N+unipredi$xgbm)/3
unipredi$predi760<-(unipredi$logi+unipredi$Red_11N+unipredi$xgbm2)/3
unipredi$predi761<-(unipredi$logi+unipredi$Red_11N+unipredi$xgbm3)/3
unipredi$predi762<-(unipredi$logi+unipredi$Red_11N+unipredi$xgbm4)/3
unipredi$predi763<-(unipredi$logi+unipredi$Red_11N+unipredi$SVM)/3
unipredi$predi764<-(unipredi$logi+unipredi$Red_11N+unipredi$SVMPoly)/3
unipredi$predi765<-(unipredi$logi+unipredi$Red_11N+unipredi$SVMRBF)/3
unipredi$predi766<-(unipredi$logi+unipredi$rf1+unipredi$gbm)/3
unipredi$predi767<-(unipredi$logi+unipredi$rf1+unipredi$xgbm)/3
unipredi$predi768<-(unipredi$logi+unipredi$rf1+unipredi$xgbm2)/3
unipredi$predi769<-(unipredi$logi+unipredi$rf1+unipredi$xgbm3)/3
unipredi$predi770<-(unipredi$logi+unipredi$rf1+unipredi$xgbm4)/3
unipredi$predi771<-(unipredi$logi+unipredi$rf1+unipredi$SVM)/3
unipredi$predi772<-(unipredi$logi+unipredi$rf1+unipredi$SVMPoly)/3
unipredi$predi773<-(unipredi$logi+unipredi$rf1+unipredi$SVMRBF)/3
unipredi$predi774<-(unipredi$logi+unipredi$rf2+unipredi$gbm)/3
unipredi$predi775<-(unipredi$logi+unipredi$rf2+unipredi$xgbm)/3
unipredi$predi776<-(unipredi$logi+unipredi$rf2+unipredi$xgbm2)/3
unipredi$predi777<-(unipredi$logi+unipredi$rf2+unipredi$xgbm3)/3
unipredi$predi778<-(unipredi$logi+unipredi$rf2+unipredi$xgbm4)/3
unipredi$predi779<-(unipredi$logi+unipredi$rf2+unipredi$SVM)/3
unipredi$predi780<-(unipredi$logi+unipredi$rf2+unipredi$SVMPoly)/3
unipredi$predi781<-(unipredi$logi+unipredi$rf2+unipredi$SVMRBF)/3
unipredi$predi782<-(unipredi$logi+unipredi$rf3+unipredi$gbm)/3
unipredi$predi783<-(unipredi$logi+unipredi$rf3+unipredi$xgbm)/3
unipredi$predi784<-(unipredi$logi+unipredi$rf3+unipredi$xgbm2)/3
unipredi$predi785<-(unipredi$logi+unipredi$rf3+unipredi$xgbm3)/3
unipredi$predi786<-(unipredi$logi+unipredi$rf3+unipredi$xgbm4)/3
unipredi$predi787<-(unipredi$logi+unipredi$rf3+unipredi$SVM)/3
unipredi$predi788<-(unipredi$logi+unipredi$rf3+unipredi$SVMPoly)/3
unipredi$predi789<-(unipredi$logi+unipredi$rf3+unipredi$SVMRBF)/3
unipredi$predi790<-(unipredi$logi+unipredi$gbm+unipredi$xgbm)/3
unipredi$predi791<-(unipredi$logi+unipredi$gbm+unipredi$xgbm2)/3
unipredi$predi792<-(unipredi$logi+unipredi$gbm+unipredi$xgbm3)/3
unipredi$predi793<-(unipredi$logi+unipredi$gbm+unipredi$xgbm4)/3
unipredi$predi794<-(unipredi$logi+unipredi$gbm+unipredi$SVM)/3
unipredi$predi795<-(unipredi$logi+unipredi$gbm+unipredi$SVMPoly)/3
unipredi$predi796<-(unipredi$logi+unipredi$gbm+unipredi$SVMRBF)/3
unipredi$predi797<-(unipredi$logi+unipredi$xgbm+unipredi$xgbm4)/3
unipredi$predi798<-(unipredi$logi+unipredi$xgbm+unipredi$SVMPoly)/3
```

```
unipredi$predi799<-(unipredi$logi+unipredi$xgbm+unipredi$SVMRBF)/3
unipredi$predi800<-(unipredi$logi+unipredi$xgbm2+unipredi$xgbm4)/3
unipredi$predi801<-(unipredi$logi+unipredi$xgbm2+unipredi$SVMPoly)/3
unipredi$predi802<-(unipredi$logi+unipredi$xgbm2+unipredi$SVMRBF)/3
unipredi$predi803<-(unipredi$logi+unipredi$xgbm3+unipredi$xgbm4)/3
unipredi$predi804<-(unipredi$logi+unipredi$xgbm3+unipredi$SVMPoly)/3
unipredi$predi805<-(unipredi$logi+unipredi$xgbm3+unipredi$SVMRBF)/3
unipredi$predi806<-(unipredi$logi+unipredi$xgbm4+unipredi$xgbm4)/3
unipredi$predi807<-(unipredi$logi+unipredi$xgbm4+unipredi$SVMPoly)/3
unipredi$predi808<-(unipredi$logi+unipredi$xgbm4+unipredi$SVMRBF)/3
unipredi$predi809<-(unipredi$rf1+unipredi$gbm+unipredi$SVM)/3
unipredi$predi810<-(unipredi$rf1+unipredi$gbm+unipredi$SVMPoly)/3
unipredi$predi811<-(unipredi$rf1+unipredi$gbm+unipredi$SVMRBF)/3
unipredi$predi812<-(unipredi$rf2+unipredi$gbm+unipredi$SVM)/3
unipredi$predi813<-(unipredi$rf2+unipredi$gbm+unipredi$SVMPoly)/3
unipredi$predi814<-(unipredi$rf2+unipredi$gbm+unipredi$SVMRBF)/3
unipredi$predi815<-(unipredi$rf3+unipredi$gbm+unipredi$SVM)/3
unipredi$predi816<-(unipredi$rf3+unipredi$gbm+unipredi$SVMPoly)/3
unipredi$predi817<-(unipredi$rf3+unipredi$gbm+unipredi$SVMRBF)/3
unipredi$predi818<-(unipredi$rf1+unipredi$xgbm+unipredi$SVM)/3
unipredi$predi819<-(unipredi$rf1+unipredi$xgbm+unipredi$SVMPoly)/3
unipredi$predi820<-(unipredi$rf1+unipredi$xgbm+unipredi$SVMRBF)/3
unipredi$predi821<-(unipredi$rf2+unipredi$xgbm+unipredi$SVM)/3
unipredi$predi822<-(unipredi$rf2+unipredi$xgbm+unipredi$SVMPoly)/3
unipredi$predi823<-(unipredi$rf2+unipredi$xgbm+unipredi$SVMRBF)/3
unipredi$predi824<-(unipredi$rf3+unipredi$xgbm+unipredi$SVM)/3
unipredi$predi825<-(unipredi$rf3+unipredi$xgbm+unipredi$SVMPoly)/3
unipredi$predi826<-(unipredi$rf3+unipredi$xgbm+unipredi$SVMRBF)/3
unipredi$predi836<-(unipredi$rf1+unipredi$xgbm2+unipredi$SVM)/3
unipredi$predi837<-(unipredi$rf1+unipredi$xgbm2+unipredi$SVMPoly)/3
unipredi$predi838<-(unipredi$rf1+unipredi$xgbm2+unipredi$SVMRBF)/3
unipredi$predi839<-(unipredi$rf2+unipredi$xgbm2+unipredi$SVM)/3
unipredi$predi840<-(unipredi$rf2+unipredi$xgbm2+unipredi$SVMPoly)/3
unipredi$predi841<-(unipredi$rf2+unipredi$xgbm2+unipredi$SVMRBF)/3
unipredi$predi842<-(unipredi$rf3+unipredi$xgbm2+unipredi$SVM)/3
unipredi$predi843<-(unipredi$rf3+unipredi$xgbm2+unipredi$SVMPoly)/3
unipredi$predi844<-(unipredi$rf3+unipredi$xgbm2+unipredi$SVMRBF)/3
unipredi$predi845<-(unipredi$rf1+unipredi$xgbm3+unipredi$SVM)/3
unipredi$predi846<-(unipredi$rf1+unipredi$xgbm3+unipredi$SVMPoly)/3
unipredi$predi847<-(unipredi$rf1+unipredi$xgbm3+unipredi$SVMRBF)/3
unipredi$predi848<-(unipredi$rf2+unipredi$xgbm3+unipredi$SVM)/3
unipredi$predi849<-(unipredi$rf2+unipredi$xgbm3+unipredi$SVMPoly)/3
unipredi$predi850<-(unipredi$rf2+unipredi$xgbm3+unipredi$SVMRBF)/3
unipredi$predi851<-(unipredi$rf3+unipredi$xgbm3+unipredi$SVM)/3
unipredi$predi852<-(unipredi$rf3+unipredi$xgbm3+unipredi$SVMPoly)/3
unipredi$predi853<-(unipredi$rf3+unipredi$xgbm3+unipredi$SVMRBF)/3
unipredi$predi854<-(unipredi$rf1+unipredi$xgbm4+unipredi$SVM)/3
unipredi$predi855<-(unipredi$rf1+unipredi$xgbm4+unipredi$SVMPoly)/3
unipredi$predi856<-(unipredi$rf1+unipredi$xgbm4+unipredi$SVMRBF)/3
unipredi$predi857<-(unipredi$rf2+unipredi$xgbm4+unipredi$SVM)/3
unipredi$predi858<-(unipredi$rf2+unipredi$xgbm4+unipredi$SVMPoly)/3
unipredi$predi859<-(unipredi$rf2+unipredi$xgbm4+unipredi$SVMRBF)/3
unipredi$predi860<-(unipredi$rf3+unipredi$xgbm4+unipredi$SVM)/3
unipredi$predi861<-(unipredi$rf3+unipredi$xgbm4+unipredi$SVMPoly)/3
unipredi$predi862<-(unipredi$rf3+unipredi$xgbm4+unipredi$SVMRBF)/3
unipredi$predi863<-(unipredi$rf1+unipredi$Red_11N+unipredi$gbm)/3
unipredi$predi864<-(unipredi$rf1+unipredi$Red_11N+unipredi$xgbm)/3
```

```
unipredi$predi865<-(unipredi$rf1+unipredi$Red_11N+unipredi$SVM)/3
unipredi$predi866<-(unipredi$rf1+unipredi$Red_11N+unipredi$SVMPoly)/3
unipredi$predi867<-(unipredi$rf1+unipredi$Red_11N+unipredi$SVMRBF)/3
unipredi$predi868<-(unipredi$Red_11N+unipredi$gbm+unipredi$SVM)/3
unipredi$predi869<-(unipredi$Red_11N+unipredi$gbm+unipredi$SVMPoly)/3
unipredi$predi870<-(unipredi$Red_11N+unipredi$gbm+unipredi$SVMRBF)/3

# 4 modelos #
unipredi$predi871<-(unipredi$logi+unipredi$rf1+unipredi$gbm+unipredi$Red_11N)/4
unipredi$predi872<-(unipredi$logi+unipredi$rf1+unipredi$xgbm+unipredi$Red_11N)/4
unipredi$predi873<-(unipredi$logi+unipredi$rf1+unipredi$xgbm+unipredi$Red_11N)/4

# 5 modelos #
unipredi$predi874<-(unipredi$logi+unipredi$rf1+unipredi$xgbm+unipredi$Red_11N+unipredi$SVM)/5
unipredi$predi875<-
(unipredi$logi+unipredi$rf1+unipredi$xgbm+unipredi$Red_11N+unipredi$SVMPoly)/5
unipredi$predi876<-
(unipredi$logi+unipredi$rf1+unipredi$xgbm+unipredi$Red_11N+unipredi$SVMRBF)/5


# Listado de modelos

dput(names(unipredi))

listado<-c("logi", "Red_3N", "Red_5N", "Red_7N", "Red_9N", "Red_11N",
        "Log_Sel", "Red_3N_Sel", "Red_5N_Sel",
        "Red_7N_Sel", "Red_9N_Sel", "Red_11N_Sel", "Red_13N_Sel",
        "bagging", "rf1", "rf2", "rf3", "gbm",
        "xgbm", "xgbm2", "xgbm3", "xgbm4", "SVM",
        "SVMPoly", "SVMRBF", "predi700", "predi701", "predi702",
        "predi703", "predi704", "predi705", "predi706", "predi707", "predi708",
        "predi709", "predi710", "predi711", "predi712", "predi713", "predi714",
        "predi715", "predi716", "predi717", "predi718", "predi719", "predi720",
        "predi721", "predi722", "predi723", "predi724", "predi725", "predi726",
        "predi727", "predi728", "predi729", "predi730", "predi731", "predi732",
        "predi733", "predi734", "predi735", "predi736", "predi737", "predi738",
        "predi739", "predi740", "predi741", "predi742", "predi743", "predi744",
        "predi745", "predi746", "predi747", "predi748", "predi749", "predi750",
        "predi751", "predi752", "predi753", "predi754", "predi755", "predi756",
        "predi757", "predi758", "predi759", "predi760", "predi761", "predi762",
        "predi763", "predi764", "predi765", "predi766", "predi767", "predi768",
        "predi769", "predi770", "predi771", "predi772", "predi773", "predi774",
        "predi775", "predi776", "predi777", "predi778", "predi779", "predi780",
        "predi781", "predi782", "predi783", "predi784", "predi785", "predi786",
        "predi787", "predi788", "predi789", "predi790", "predi791", "predi792",
        "predi793", "predi794", "predi795", "predi796", "predi797", "predi798",
        "predi799", "predi800", "predi801", "predi802", "predi803", "predi804",
        "predi805", "predi806", "predi807", "predi808", "predi809", "predi810",
        "predi811", "predi812", "predi813", "predi814", "predi815", "predi816",
        "predi817", "predi818", "predi819", "predi820", "predi821", "predi822",
        "predi823", "predi824", "predi825", "predi826", "predi836", "predi837",
        "predi838", "predi839", "predi840", "predi841", "predi842", "predi843",
        "predi844", "predi845", "predi846", "predi847", "predi848", "predi849",
        "predi850", "predi851", "predi852", "predi853", "predi854", "predi855",
        "predi856", "predi857", "predi858", "predi859", "predi860", "predi861",
        "predi862", "predi863", "predi864", "predi865", "predi866", "predi867",
        "predi868", "predi869", "predi870", "predi871", "predi872", "predi873",
```

```
        "predi874", "predi875", "predi876")

# Defino funcion tasafallos

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}

auc<-function(x,y) {
 curvaroc<-roc(response=x,predictor=y)
 auc<-curvaroc$auc
 return(auc)
}

# Se obtiene el numero de repeticiones CV y se calculan las medias por repe en
# el data frame medias0

repeticiones<-nlevels(factor(unipredi$Rep))
unipredi$Rep<-as.factor(unipredi$Rep)
unipredi$Rep<-as.numeric(unipredi$Rep)


medias0<-data.frame(c())
for (prediccion in listado)
{
 unipredi$proba<-unipredi[,prediccion]
 unipredi[,prediccion]<-ifelse(unipredi[,prediccion]>0.5,"Yes","No")
 for (repe in 1:repeticiones)
 {
  paso <- unipredi[(unipredi$Rep==repe),]
  pre<-factor(paso[,prediccion])
  archi<-paso[,c("proba","obs")]
  archi<-archi[order(archi$proba),]
  obs<-paso[,c("obs")]
  tasa=1-tasafallos(pre,obs)
  t<-as.data.frame(tasa)
  t$modelo<-prediccion
  auc<-auc(archi$obs,archi$proba)
  t$auc<-auc
  medias0<-rbind(medias0,t)
 }
}

# Finalmente boxplot

par(cex.axis=0.5,las=2)
boxplot(data=medias0,tasa~modelo,col="grey",main="TASA FALLOS")

# Para AUC se utiliza la variable auc del archivo medias0

boxplot(data=medias0,auc~modelo,col="grey",main="AUC")

# PRESENTACION TABLA MEDIAS

tablamedias<-medias0 %>%
```

```
 group_by(modelo) %>%
 summarize(tasa=mean(tasa))

tablamedias<-tablamedias[order(tablamedias$tasa),]
head(tablamedias,10)
#   modelo    tasa
# 1 predi703 0.147
# 2 predi836 0.147
# 3 predi783 0.147
# 4 predi775 0.147
# 5 predi776 0.147
# 6 predi768 0.147
# 7 predi839 0.147
# 8 predi842 0.148
# 9 predi821 0.148
#10 predi784 0.148

tail(tablamedias,10)
#   modelo      tasa
# 1 Red_13N_Sel 0.160
# 2 Red_11N     0.160
# 3 Red_9N_Sel  0.161
# 4 Red_3N_Sel  0.162
# 5 Red_3N      0.162
# 6 Red_9N      0.163
# 7 Red_7N      0.163
# 8 Red_7N_Sel  0.164
# 9 Red_5N_Sel  0.164
# 10 Red_5N     0.166



# ORDENACIÓN DEL FACTOR MODELO POR LAS MEDIAS EN TASA
# PARA EL GRAFICO

medias0$modelo <- with(medias0,
                reorder(modelo,tasa, mean))
par(cex.axis=0.7,las=2)
boxplot(data=medias0,tasa~modelo,col="grey", main='TASA FALLOS')

# ***********************************
# PARA AUC
# ***********************************

# PRESENTACION TABLA MEDIAS

tablamedias2<-medias0 %>%
 group_by(modelo) %>%
 summarize(auc=mean(auc))

tablamedias2<-tablamedias2[order(-tablamedias2$auc),]

head(tablamedias2,10)
# modelo     auc
# 1 predi790 0.910
# 2 predi791 0.910
# 3 predi706 0.910
```

```
# 4 predi797 0.910
# 5 predi792 0.910
# 6 predi707 0.910
# 7 predi800 0.909
# 8 predi775 0.909
# 9 predi767 0.909
#10 predi793 0.909

tail(tablamedias2,10)
# modelo       auc
# 1 bagging    0.894
# 2 SVMPoly    0.894
# 3 Red_11N    0.894
# 4 Red_9N     0.892
# 5 Red_7N_Sel 0.891
# 6 Red_7N     0.890
# 7 Red_5N_Sel 0.887
# 8 Red_3N_Sel 0.885
# 9 Red_5N     0.884
#10 Red_3N     0.880

# ORDENACIÓN DEL FACTOR MODELO POR LAS MEDIAS EN AUC
# PARA EL GRAFICO

medias0$modelo <- with(medias0,
                  reorder(modelo,auc, mean))
par(cex.axis=0.7,las=2)
boxplot(data=medias0,auc~modelo,col="grey", main='AUC')



# Se pueden escoger listas pero el factor hay que pasarlo a character
# para que no salgan en el boxplot todos los niveles del factor

listadobis<-c("logi", "Red_3N", "Red_5N", "Red_7N", "Red_9N", "Red_11N",
          "Log_Sel", "Red_3N_Sel", "Red_5N_Sel",
          "Red_7N_Sel", "Red_9N_Sel", "Red_11N_Sel", "Red_13N_Sel",
          "bagging", "rf1", "rf2", "rf3", "gbm",
          "xgbm", "xgbm2", "xgbm3", "xgbm4", "SVM",
          "predi703", # Logistica + RandomForest2
          "predi706", # Logistica + XGBOOST
          "predi775", # Logistica + RandomForest2 + XGBOOST
          "predi763", # Logistoca + Red 11 Nodos 0,01 LR + SVMLIneal
          "predi790", # Logistica + GBM + XGBOOST
          "predi800", # Logistica+XGBOOST2+XGBOOST4
          "predi836"  # RandomForest1 + XGBOOST2 + SVM
)

medias0$modelo<-as.character(medias0$modelo)

mediasver<-medias0[medias0$modelo %in% listadobis,]

mediasver$modelo <- with(mediasver,
                   reorder(modelo,tasa, median))

par(cex.axis=0.9,las=2)
boxplot(data=mediasver,tasa~modelo,col="grey",main='TASA FALLOS')
```

```
mediasver$modelo <- with(mediasver,
                  reorder(modelo,auc, median))

par(cex.axis=0.9,las=2)
boxplot(data=mediasver,auc~modelo,col="grey",main='AUC')


# Se pueden escoger listas pero el factor hay que pasarlo a character
# para que no salgan en el boxplot todos los niveles del factor

listadobis2<-c("logi","gbm",
          "xgbm", "xgbm2", "xgbm3", "xgbm4",
          "predi703", # Logistica + RandomForest2
          "predi706", # Logistica + XGBOOST
          "predi775", # Logistica + RandomForest2 + XGBOOST
          "predi763", # Logistoca + Red 11 Nodos 0,01 LR + SVMLIneal
          "predi790", # Logistica + GBM + XGBOOST
          "predi800", # Logistica+XGBOOST2+XGBOOST4
          "predi836"  # RandomForest1 + XGBOOST2 + SVM
)

medias0$modelo<-as.character(medias0$modelo)

mediasver2<-medias0[medias0$modelo %in% listadobis2,]

mediasver2$modelo <- with(mediasver2,
                  reorder(modelo,tasa, median))

#-------------------------------------------------------------------------------#
## COMPARACION DE RESULTADOS R y SAS ##
# Se carga el archivo de modelos de SAS
path_sas='D:/Fede/Google Drive/Master Minería de Datos e Inteligencia de Negocios/Materias/Tecnicas
de Machine Learning/Trabajos/Clasificación/SAS/'

sas <- read.xlsx(paste0(path_sas,'tasafallos_bestmodels.xlsx'),colNames=TRUE)
sas$tasa<-as.numeric(sas$tasa) # transformo la tasa a número

mediasver3<-rbind(mediasver2,sas)

mediasver3$origen<-as.factor(mediasver3$origen)

mediasver3$modelo <- with(mediasver3,
                   reorder(modelo,tasa, median))

par(cex.axis=0.9,las=2)
boxplot(data=mediasver3,
       tasa~modelo,
       col=mycolors,
       main='TASA FALLOS')

legend("topleft",
      legend=c("SAS","R"),
      col=c("blue","pink"),
      pch=16)
```

```
listadobis3<-c("logi","gbm",
           "xgbm", "xgbm2", "xgbm3", "xgbm4",
           "predi703", # Logistica + RandomForest2
           "predi706", # Logistica + XGBOOST
           "predi775", # Logistica + RandomForest2 + XGBOOST
           "predi763", # Logistoca + Red 11 Nodos 0,01 LR + SVMLIneal
           "predi790", # Logistica + GBM + XGBOOST
           "predi800", # Logistica+XGBOOST2+XGBOOST4
           "predi836",  # RandomForest1 + XGBOOST2 + SVM
           "RED",
           "RLOG",
           "REDBOO",
           "LBOOST",
           "R-L-BOO")

mediasver4<-mediasver3[mediasver3$modelo %in% listadobis3,]

mediasver4$modelo <- with(mediasver4,
                  reorder(modelo,tasa, median))

par(cex.axis=0.9,las=2)
boxplot(data=mediasver4,
      tasa~modelo,
      col=c("red","red","red","red","red","blue","blue","red","red","red","red","red","red","red","red",
          "blue","blue","red","blue","blue"),
      main='TASA FALLOS')

legend("topleft",
     legend=c("SAS","R"),
     col=c("blue","red"),
     cex=1.5,
     pch=16)

# GRÁFICOS DE APOYO PARA OBSERVAR COMPORTAMIENTO DE LOS MODELOS

unipredi<-cbind(predi1,predi2,predi3,predi4,predi5,predi6,predi7,
           predi11,predi21,predi22,predi23,predi24,predi25,predi26,
           predi301,predi302,predi303,predi304,predi305,predi306,
           predi401,
           predi501,predi502,predi503,predi504,predi505,
           predi601,predi602,predi603)
# Esto es para eliminar columnas duplicadas
unipredi<- unipredi[, !duplicated(colnames(unipredi))]
# Añadir ensamblados
unipredi$predi703<-(unipredi$logi+unipredi$rf2)/2
unipredi$predi706<-(unipredi$logi+unipredi$xgbm)/2
unipredi$predi763<-(unipredi$logi+unipredi$Red_11N+unipredi$SVM)/3
unipredi$predi775<-(unipredi$logi+unipredi$rf2+unipredi$xgbm)/3
unipredi$predi790<-(unipredi$logi+unipredi$gbm+unipredi$xgbm)/3
unipredi$predi800<-(unipredi$logi+unipredi$xgbm2+unipredi$xgbm4)/3
unipredi$predi836<-(unipredi$rf1+unipredi$xgbm2+unipredi$SVM)/3


#-------------------------------------------------------------------------------------------------#
# Guardo las prediciones en un excel
library(writexl)
write_xlsx(unipredi, paste0(path,'Telco_customer_Churn_Predictions.xlsx'))
```

```
#-------------------------------------------------------------------------------------------------#
# Confusion matrix

# logi
confusionMatrix(medias1[[2]][["pred"]],medias1[[2]][["obs"]],"Yes")

# log_Sel
confusionMatrix(medias11[[2]][["pred"]],medias11[[2]][["obs"]],"Yes")

# Red 11 nodos
confusionMatrix(medias6[[2]][["pred"]],medias6[[2]][["obs"]],"Yes")

# rf4
confusionMatrix(medias306[[2]][["pred"]],medias306[[2]][["obs"]],"Yes")

# gbm
confusionMatrix(medias401[[2]][["pred"]],medias401[[2]][["obs"]],"Yes")

# xgbm
confusionMatrix(medias502[[2]][["pred"]],medias502[[2]][["obs"]],"Yes")

# SVM lineal
confusionMatrix(medias601[[2]][["pred"]],medias601[[2]][["obs"]],"Yes")


#-------------------------------------------------------------------------------------------------#
# importancia de variables

data11 <- data1
data11$Partner <- NULL

# logistica
set.seed(12345)

control<-trainControl(method = "repeatedcv",number=10,repeats=10,
                savePredictions = "all",classProbs=TRUE)

logistic_var <- train(factor(Churn)~.,data=data11,
                trControl=control,method="glm",family = binomial(link="logit"))

summary(logistic_var)

logistic_var$modelInfo$parameters
exp(coef(logistic_var$finalModel))

# Random forest (rf4)
set.seed(12345)
rfgrid<-expand.grid(mtry=20)

control<-trainControl(method = "repeatedcv",number=10,repeats=10,
                savePredictions = "all", classProbs=TRUE)

rf_var <- train(factor(Churn)~.,data=data11,
        method="rf",trControl=control,tuneGrid=rfgrid,
        linout = FALSE,ntree=600,nodesize=20,replace=TRUE,
        importance=TRUE)
```

```
final<-rf_var$finalModel

tabla<-as.data.frame(importance(final))
tabla<-tabla[order(-tabla$MeanDecreaseAccuracy),]
tabla
# MeanDecreaseAccuracy indica lo que baja la precision (tasa de aciertos)
# en caso de no contar con esa variable en el modelo

barplot(tabla$MeanDecreaseAccuracy,names.arg=rownames(tabla))

# GBM
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.04),
                n.minobsinnode=c(30),
                n.trees=c(500),
                interaction.depth=c(2))

control<-trainControl(method = "repeatedcv",number=10,repeats=10,savePredictions = "all",
                 classProbs=TRUE)

gbm_var <- train(factor(Churn)~.,data=data11,
         method="gbm",trControl=control,tuneGrid=gbmgrid,
         distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm_var
summary(gbm_var)

tabla<-summary(gbm_var)
par(cex=0.8,las=2)
barplot(tabla$rel.inf,names.arg=row.names(tabla))

# XG BOOST (xgbm2)

set.seed(12345)

xgbmgrid<-expand.grid(eta=c(0.01),
                 min_child_weight=c(10),
                 nrounds=c(2500),
                 max_depth=6,
                 gamma=0,
                 colsample_bytree=1,subsample=1)

control<-trainControl(method = "repeatedcv",number=10,repeats=10,savePredictions = "all",
                 classProbs=TRUE)


xgbm_var<- train(factor(Churn)~.,data=data11,
           method="xgbTree",trControl=control,
           tuneGrid=xgbmgrid,objective = "binary:logistic",verbose=FALSE,
           alpha=3,lambda=6,lambda_bias=10)

xgbm_var

varImp(xgbm_var)
varImp(xgbm_var, scale=FALSE)
```

```
#-------------------------------------------------------------------------------------------------#
## Segmentación ##

var_clu<-c("Customer_ID", "Churn", "CLTV","Tenure_in_Months",
        "Contract_Month_to_Month", "Contract_One_Year", "Contract_Two_Year",
        "Number_of_Referrals_0", "Number_of_Referrals_1",
        "Number_of_Referrals_2", "Number_of_Referrals_3", "Number_of_Referrals_4",
        "Number_of_Referrals_5", "Number_of_Referrals_6", "Number_of_Referrals_7",
        "Number_of_Referrals_8", "Number_of_Referrals_9", "Number_of_Referrals_10",
        "Number_of_Dependents_0", "Number_of_Dependents_1", "Number_of_Dependents_2",
        "Number_of_Dependents_3",
        "Age", "Population",
        "Avg_Monthly_Long_Distance_Charge",
        "Avg_Monthly_GB_Download", "Monthly_Charge", "Total_Charges",
        "Total_Extra_Data_Charges", "Total_Long_Distance_Charges", "Total_Refunds"
) # 31

databisclu <- databis[,var_clu]

dput(colnames(databisclu[,3:31])) # reviso las variables que deseo usar para la segmentacion

## Buscar el K optimo ##
# Metodos

library(factoextra)
library(NbClust)

set.seed(123)

# Elbow method
elbow <- fviz_nbclust(databisclu[3:31], kmeans, nstart = 25, method = "wss", nboot = 50) +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")
#

# Silhouette method
silhouette <- fviz_nbclust(databisclu[3:31], kmeans, nstart = 25, method = "silhouette", nboot = 50) +
  labs(subtitle = "Silhouette method")

# Gap statistic
set.seed(123)
gap <- fviz_nbclust(databisclu[3:31], kmeans, nstart = 25,  method = "gap_stat", nboot = 50) +
  labs(subtitle = "Gap statistic method")


# Se genera la distancia en forma de matriz
diss_matrix<- dist(databisclu[3:31], method = "euclidean", diag=FALSE)

# se prueban diferentes métodos
res <- NbClust(data = databisclu[3:31], diss=diss_matrix, distance = NULL, min.nc=3, max.nc=10,
           method = "kmeans", index="silhouette")
res$Best.nc
# conviene 4 clusters

res1 <- NbClust(data = databisclu[3:31], diss=diss_matrix, distance = NULL, min.nc=3, max.nc=10,
```

```
            method = "kmeans", index="hubert")
# conviene 4 clusters
res1$Best.nc


## internal validation
library(clValid)
library(kohonen)
library(mclust)

intern <- clValid(databisclu[3:31], nClust = 3:8,
              clMethods = c("kmeans"),
              validation = c("internal"),
              maxitems=nrow(databisclu[3:31]))

optimalScores(intern)
plot(intern)

par(mfrow=c(1,1))

#-------------------------------------------------------------------------------------------------------------#

# K-Means #

set.seed(12345)

km <- kmeans(databisclu[,3:31], 4, nstart = 25)
# cluster por las variables seleccionadas

# Resultados
print(km)

# Asignación de cluster por observacion
km$cluster

# Cantidad de observaciones por cluster
km$size

# centroides
km$centers

#-------------------------------------------------------------------------------------------------------------#
# Se incorpora el número de cluster a cada una de las observaciones
data2 <- cbind(data,km$cluster)

# se renombra la columna
names(data2)[names(data2) == "km$cluster"] <- "cluster"

# se carga el excel original
excel1 <- read.xlsx(paste0(path,'Telco_customer_churn.xlsx'),colNames=TRUE)

# se une el cluster al excel
excel2 <-cbind(excel1, data2$cluster)

names(excel2)[names(excel2) == "data2$cluster"] <- "Cluster"

# se reemplazan los nombres con "." por "_"
```

```
colnames(excel2)

colnames(excel2) <- gsub('[.]', '_', colnames(excel2))

# se guarda el excel resultante
library(writexl)
write_xlsx(excel2, paste0(path,'Telco_customer_churn_Cluster.xlsx'))
```